DOCUMENT RESUME

ED 326 649 CE 056 395

TITLE Military Training. Its Effectiveness for Technical Specialties Is Unknown. Report to the Secretary of Defense.

INSTITUTION General Accounting Office, Washington, DC. Program Evaluation and Methodology Div.

REPORT NO GAO/PEMD-91-4
PUB DATE Oct 90
NOTE 106p.
AVAILABLE FROM U.S. General Accounting Office, P.O. Box 6015, Gaithersburg, MD 20877 (first five copies free; additional copies $2.00 each; 100 or more--25% discount).
PUB TYPE Reports - Evaluative/Feasibility (142) --

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Armed Forces; Data Collection; *Evaluation Criteria; *Evaluation Methods; Evaluation Needs; *Evaluation Problems; Evaluation Utilization; Females; Job Performance; *Military Training; Minority Groups; *Outcomes of Education; Postsecondary Education; *Program Effectiveness; Program Evaluation; Reliability; Technical Education; Validity

ABSTRACT

A study examined the information collected by the Department of Defense on both the quality of its new recruits and the effectiveness of its training in preparing recruits to operate in a technologically sophisticated environment. It found that data were collected at a recruit's entrance to military life, during and upon completion of formal training, and after assignment to a military specialty in the field. The study showed that for most recruits, the services' selection criteria are moderately successful at predicting individual performance during classroom technical training. However, they are notably less successful for women and minority recruits. Each service has evaluation mechanisms in place, but only the Army systematically collects data on the field performance of individual graduates in a way that would allow comparison of a graduate's on-the-job performance with entry-level ability and classroom performance. These data reveal an even weaker connection for women and minority group members between criteria used to assign them to technical specialties and their later field performance. The field evaluation practices of the Navy are particularly fragmented and have deteriorated during the 1980s. The lack of reliable field performance data in the Navy and the Air Force makes realistic assessment of training effectiveness impossible. The study also found that the quality of recruits in respect to education increased during the early 1980s, but leveled off later in the decade. (KC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

United States General Accounting Office

GAO

Report to the Secretary of Defense

October 1990

MILITARY TRAINING

Its Effectiveness for Technical Specialties Is Unknown

ED326649

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

GAO/PEMD-91-4

2

GAO

United States
General Accounting Office
Washington, D.C. 20548

Program Evaluation and
Methodology Division

B-239914

October 16, 1990

The Honorable Richard B. Cheney
The Secretary of Defense

Dear Mr. Secretary:

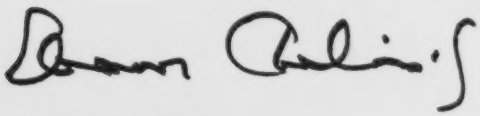
In this report, we review the information sources on which the services base their evaluations of the effectiveness of their technical training programs, recruit selection, and classification decisions. We undertook this review because the technical sophistication of modern weaponry has intensified the need for well-qualified recruits and effective technical training. This report identifies some critical gaps in the services' ability to measure how effectively they are selecting and preparing recruits to use and maintain today's complex weapons systems.

This report contains recommendations in Chapter 5. The head of a federal agency is required by 31 U.S.C. 720 to submit a written statement on actions taken on these recommendations to the Senate Committee on Governmental Affairs and the House Committee on Government Operations not later than 60 days after the date of the report and to the House and Senate Committees on Appropriations with the agency's first request for appropriations made more than 60 days after the date of the report.

We are sending copies of this report to appropriate House and Senate committees, members of Congress from the states mentioned in the report, and the Director of the Office of Management and Budget. We will also make copies available to interested organizations, as appropriate, and to others upon request.

If you have any questions or would like additional information, please call me at (202) 275-1854. Major contributors to the report are listed in appendix VI.

Sincerely yours,



Eleanor Chelimsky
Assistant Comptroller General

Executive Summary

Purpose

The ability of the armed forces to carry out their mission into the next century will depend on both hardware and personnel considerations: the reliability and appropriateness of weapons systems, the quality of military personnel, and the "fit" of human skills to the operating demands of weapons systems. If the entry-level aptitude, knowledge, and skills of new recruits should fall short of the human requirements needed to operate and maintain new technologically sophisticated systems, greater demands would be placed on the armed services to compensate for the shortfall through training. The purpose of this report was to examine the information collected by the Department of Defense (DOD) on both the quality of its new recruits and the effectiveness of its training in preparing recruits to operate in a technologically sophisticated environment.

Background

A recruit is admitted to military service and assigned to an occupational specialty on the basis of tests taken at recruitment. Upon completion of basic training, most recruits receive additional classroom training in their specialty and then are assigned to perform the specialty in the field. This typical sequence encompasses the three points in a recruit's service career where data critical to evaluating the success of training must be collected: at entrance to military life, during and upon completion of formal training, and after assignment to a military specialty in the field.

An adequate system of assessing training effectiveness must include reliable and valid information at each of these points, and should examine the interrelationships among these data points to test the congruence of initial selection and placement data, classroom measures, and the ultimate criterion—field performance.

During the mid-1980's, the services reported dramatic improvements in the general qualifications of new recruits. The improvements were attributed to better compensation and educational benefits, increased recruiting efforts, and heightened public appreciation of the military role. These reports did not, however, address the specific area of technical qualifications among recruits. More recently, the services have reported difficulty in filling their quotas with highly qualified recruits. This perceived decline in the ability levels of recruits entering training raises questions about the reality of that decline, about its magnitude, about the effectiveness of the process by which recruits are selected for training, and about the actual on-the-job performance of those recruits.

Results in Brief

GAO found that the aptitude level of recruits did increase during the 1980's but that most of the improvement occurred during the first half of the decade. Since then, little change has occurred in general aptitude for training, but the levels of some of the more technical skills have declined among recruits, in one case below the 1981 level. Women and members of minority groups consistently scored lower in tests used to assign recruits to more technical occupational specialties such as radar specialist positions.

GAO concluded that, for most recruits, the services' selection criteria are moderately successful at predicting individual performance during classroom technical training. However, they are notably less successful for women and minority recruits.

Each service has evaluation mechanisms in place, but only the Army systematically collects data on the field performance of individual graduates in a way that would allow comparison of a graduate's on-the-job performance with his or her entry-level ability and classroom performance. These data reveal an even weaker connection for women and minority group members between criteria used to assign them to technical specialties and their later field performance. The field evaluation practices of the Navy are particularly fragmented and have deteriorated during the 1980's. GAO found that the lack of reliable field performance data in the Navy and the Air Force makes realistic assessment of training effectiveness impossible.

GAO concluded that the insensitivity of selection and placement measures as predictors of future success for female and minority recruits is a matter of serious concern in view of the military's increasing reliance on these groups to perform technical roles.

Principal Findings

Recent Quality Trends

All services administer the Armed Services Vocational Aptitude Battery (ASVAB) to new recruits. The primary measure of a recruit's aptitude is the Armed Forces Qualification Test (AFQT), which is made up of four ASVAB subtests. AFQT scores have tended to level off after rising in the early 1980's. Average scores on three of the four subtests used to select candidates for technical training have declined since mid-decade, and scores on one—the Electronics Information subtest—are lower than in

1981. A smaller percentage of recruits now qualify for the most demanding technical specialties than at any time since 1981. Women and minority group members are severely underrepresented among qualifiers because they score lower, on average, than white males. (See pages 18-31.)

Classroom Evaluation Measures

Each service has established evaluation mechanisms to monitor instructional quality and curriculum coverage in classroom training. Overall, the grading procedures in the courses GAO reviewed appeared to discriminate acceptably well among levels of student performance (with the exception of some Army courses where recorded grades were unreliable indicators of classroom performance). (See pages 32-34, 36-38, and 40-41.)

Selection criteria from ASVAB are moderately successful in predicting the performance of most students for training, but are significantly less reliable predictors for women and minority students. While these groups appeared to overcome their lower scores on aptitude measures in the Navy and Air Force courses reviewed, the differences in classroom performance for nonwhite and female students persisted throughout the Army technical courses reviewed. (See pages 34-36, 38-39, and 40-41.)

GAO developed a statistically more sophisticated summary score from ASVAB using factor analysis. This factor score generally performed better than AFQT and the Electronics Composite score in predicting final grades for all demographic groupings. This finding suggests that broader-based selection criteria than those currently in use could be more reliable predictors of classroom performance, at least in the technical areas GAO reviewed. (See pages 36, 39, and 41.)

Field Measures of Training Effectiveness

The Army's Skill Qualification Test provides the only objective, systematically collected estimates of the field performance of individual graduates of training. The Air Force and the Navy rely instead largely on feedback mechanisms through which field commanders and supervisors may submit complaints to the training community if they believe their graduates have been inadequately trained. In addition, Air Force evaluation units periodically survey a sample of supervisors of course graduates for their perceptions of the quality and appropriateness of training. A similar practice was followed in the Navy until the mid-1980's. Internal reports have been sharply critical of the quality of the Navy's

training assessment procedures, but these deficiencies are only slowly being corrected. (See pages 45-50.)

Field performance measures have been developed by DOD under the Joint-Service Job Performance Measurement project and may be applicable to training assessment purposes. (See page 51.)

ASVAB scores in our sample are weaker predictors of field performance as measured by the Army than they are of classroom performance and only predict well for white male recruits. The factor scores developed by GAO are better predictors than either AFQT or the Electronics qualifying scores used by the Army. No ASVAB score was significantly correlated with field performance for women or minority soldiers. (See pages 45-46.)

Recommendations

GAO believes that evaluating the effectiveness of the training provided by the services is crucial if they are to meet the future challenges of changing demographics and increasingly sophisticated weaponry. GAO therefore recommends that the Assistant Secretary of Defense for Force Management and Personnel attempt to develop more sensitive indicators of classroom and field performance in technical specialties for women and minority recruits from extant data. GAO also recommends that the Assistant Secretary review alternative measures of field performance already developed by the services under the Job Performance Measurement project for their applicability to training and on-the-job performance evaluation. GAO further recommends that the Secretary of the Army direct the Training and Doctrine Command to review for accuracy, appropriateness, and reliability the classroom grading procedures identified within the report as deficient. Finally, GAO recommends that the Secretary of the Navy establish a firm deadline for developing a training evaluation program and that he direct that current resources allocated to this effort be reexamined for their adequacy.

Agency Comments

In a written response to a draft of this report, DOD concurred with all of its recommendations and identified specific actions to be taken toward implementing them. DOD also concurred or partially concurred with what it identified as the main findings contained in the report. (See appendix V.) We have reviewed these comments and, where appropriate, have made changes to the text.

Contents

Executive Summary		2
Chapter 1 Introduction		10
	Recruit Quality in the 1980's	10
	Recruit Training	11
	Objectives, Scope, and Methodology	13
	Strengths and Limitations of Our Study	17
Chapter 2 The Quality of Military Recruits: 1981-89		18
	Armed Services Vocational Aptitude Battery (ASVAB)	18
	Summary and Conclusions	30
Chapter 3 Classroom Measures of Training Effectiveness		32
	Army	33
	Navy	36
	Air Force	39
	Summary and Conclusions	42
Chapter 4 Field Measures of Training Effectiveness		45
	Army	45
	Navy	48
	Air Force	50
	Alternative Data Sources: The Job Performance Measurement Project	51
	Summary and Conclusions	52
Chapter 5 Summary, Recommendations, and Agency Comments and Our Response		53
	Summary	53
	Recommendations	54
	Agency Comments and Our Response	55

Appendixes

Appendix I: AFQT Mean Score and Electronics Composite Summary Statistics: 1981-89 60
Appendix II: Predictor and Criterion Variable Mean Scores 64
Appendix III: Intercorrelation of Study Variables by Occupational Specialty 66
Appendix IV: Army SQT Mean Scores, by Occupational Specialty 77
Appendix V: Comments From the Department of Defense 78
Appendix VI: Major Contributors to This Report 103

Tables

Table 1.1: How AFQT Test Results Are Categorized 15
Table 3.1: Army Occupational Specialties Reviewed 33
Table 3.2: Mean Scores on Predictor and Criterion Variables, Army 34
Table 3.3: Intercorrelation of Study Variables, Army 35
Table 3.4: Occupational Specialties Reviewed, Navy 37
Table 3.5: Mean Scores on Predictor and Criterion Variables, Navy 37
Table 3.6: Intercorrelation of Study Variables, Navy 39
Table 3.7: Occupational Specialties Reviewed, Air Force 40
Table 3.8: Mean Scores on Predictor and Criterion Variables, Air Force 40
Table 3.9: Intercorrelation of Study Variables, Air Force 42
Table 4.1: Correlation of SQT and Predictor Variables 46
Table I.1: AFQT Mean Scores, by Gender 60
Table I.2: AFQT Mean Scores, by Service 60
Table I.3: AFQT Mean Scores, by Race/Ethnicity 61
Table I.4: AFQT Mean Score Overall Totals 61
Table I.5: Electronics Composite Mean Scores, by Gender 62
Table I.6: Electronics Composite Mean Scores, by Service 62
Table I.7: Electronics Composite Mean Scores, by Race/Ethnicity 63
Table I.8: Electronics Composite Mean Score Overall Totals 63
Table II.1: Army Mean Scores 64
Table II.2: Navy Mean Scores 64
Table II.3: Air Force Mean Scores 65
Table III.1: Intercorrelation of Study Variables: Army, 24J 66

Table III.2: Intercorrelation of Study Variables: Army, 27N 67
Table III.3: Intercorrelation of Study Variables: Army, 29V 68
Table III.4: Intercorrelation of Study Variables: Navy, AQ 69
Table III.5: Intercorrelation of Study Variables: Navy, AX 70
Table III.6: Intercorrelation of Study Variables: Navy, STG 71
Table III.7: Intercorrelation of Study Variables: Navy, STS 72
Table III.8: Intercorrelation of Study Variables: Air Force, 45530A 73
Table III.9: Intercorrelation of Study Variables: Air Force, 45530B 74
Table III.10: Intercorrelation of Study Variables: Air Force, 30332 75
Table III.11: Intercorrelation of Study Variables: Air Force, 30333 76

Figures

Figure 1.1: Recruit Training Process 12
Figure 1.2: Data Sources and Comparisons 14
Figure 2.1: Mean AFQT Scores, by Gender: 1981-89 19
Figure 2.2: Mean AFQT Scores, by Race/ Ethnicity: 1981-89 20
Figure 2.3: Mean AFQT Scores, by Service: 1981-89 21
Figure 2.4: Mean AFQT Subtest Scores, 1981-89 22
Figure 2.5: Mean Electronics Composite Scores, by Gender: 1981-89 23
Figure 2.6: Mean Electronics Composite Scores, by Race/ Ethnicity: 1981-89 24
Figure 2.7: Mean Electronics Composite Scores, by Service: 1981-89 25
Figure 2.8: Mean Electronics Composite Subtest Scores, 1981-89 26
Figure 2.9: Number of Recruits Qualifying for Training as Control and Warning Radar Specialists, 1981-89 27
Figure 2.10: Percent of Recruits Qualifying for Training as Control and Warning Radar Specialists, 1981-89 28
Figure 2.11: Number of Recruits Qualifying for Training as Systems Repair Technicians, 1981-89 29
Figure 2.12: Percent of Recruits Qualifying for Training as Systems Repair Technicians, 1981-89 30

Abbreviations

AFQT	Armed Forces Qualification Test
ASVAB	Armed Services Vocational Aptitude Battery
DOD	Department of Defense
FLETAP	Fleet Training Assessment Program
GAO	General Accounting Office
ISD	Instructional System Development
JPM	Job Performance Measurement
NTSC	Naval Training Systems Center
SQT	Skill Qualification Test
TAST	Training Assessment Survey Team

Introduction

The ability of the armed forces to carry out their mission into the next century will depend on both hardware and personnel considerations: the reliability and appropriateness of weapons systems, the quality of military personnel, and the "fit" of human skills to the operating demands of weapons systems. If the entry level aptitude, knowledge, and skills of new recruits should fall short of the human requirements needed to operate and maintain new technologically sophisticated weapons systems, greater demands would be placed on the armed services to compensate for the shortfall through training. In this report, we will examine the information collected by DOD on both the quality of its new recruits and the effectiveness of its training in preparing recruits to operate in a technologically sophisticated military environment.

Recruit Quality in the 1980's

In hearings before the House Appropriations Committee on the fiscal year 1988 budget for DOD, the Assistant Secretary for Force Management and Personnel characterized the changes since 1980 in the nation's armed forces in these words: "Today we are recruiting the highest quality personnel in history. [The services' personnel possess]. . . high intelligence, correct experience mix, [and] high skill levels." The reasons cited for this "most remarkable turnaround in peacetime history" were many: higher pay and improved quality of life for members of the armed forces; the recession and consequent unemployment of the early 1980's, which widened the pool of applicants; improved educational benefits for military service; more intensive and effective recruiting; and recovery from the poor public perception of the military following the war in Vietnam.

The statistics cited by DOD supported this favorable view. In 1980, 68 percent of recruits were high school graduates (versus 75 percent for the youth population in general). By 1986, 92 percent of recruits had high school diplomas. Whereas 65 percent of recruits in 1980 scored in the top three mental categories on the Armed Forces Qualification Test (versus 69 percent for the norm group), in 1986, 96 percent achieved this level.

Yet the demographic and educational realities of the immediate future are likely to affect this optimistic scenario. The number of young people available for the military recruit pool will continue to diminish until the

mid-1990's.[1] The composition of the recruit pool will also shift. According to research sponsored by the Department of Labor, by the year 2000 five of every six new labor force entrants will be female, minority group members, or immigrants.[2] Meanwhile, the graduates of the American educational system are said to be falling further behind the youth of competitor nations in technological literacy at the same time that U.S. weapons systems are becoming increasingly sophisticated.[3]

DOD has also begun to voice concern. Hints of uneasiness emerged in the fiscal year 1988 appropriations hearings when the Air Force reported increased difficulty in securing quality recruits. In the same hearings, the Navy expressed its concern over the steady erosion of its Delayed Entry Pool—the program under which applicants agree to enter the service within a year. In addition, for the first time in eight years, the Army failed to meet its quarterly recruiting quota in the first quarter of fiscal year 1989.

Recruit Training

Figure 1.1 identifies the typical sequence that occurs during the early stages of a recruit's time in the military. As shown, after their basic training—the length and content of which varies by service—most recruits attend additional training to equip them to function effectively in some occupational specialty. The recruit's area of specialization is determined by service needs, qualifications as determined on tests administered during the recruiting process, and individual interests.

[1]U.S. Bureau of the Census, Projections of the Population of the United States, by Age, Sex, and Race: 1988 to 2080. Current Population Reports, Series P-25, No. 1018 (Washington, D.C.: U.S. Government Printing Office, 1989), p. 6.

[2]William B. Johnston and Arnold H. Packer, Workforce 2000: Work and Workers for the 21st Century (Indianapolis, Indiana: Hudson Institute, 1987), p.95. See also U.S. Office of Personnel Management, Civil Service 2000 (Washington, D.C.: U.S. Government Printing Office, 1988).

[3]Martin Binkin, Military Technology and Defense Manpower (Washington, D.C.: The Brookings Institution, 1986). See also Aerospace Education Foundation, America's Next Crisis: The Shortfall in Technical Manpower (Arlington, Va.: The Aerospace Education Foundation, 1989); and National Research Council, A Challenge in Numbers: People in the Mathematical Sciences (Washington, D.C.: National Academy of Sciences, 1990).

Figure 1.1: Recruit Training Process

The training curriculum for each occupational specialty is designed through a structured set of procedures called Instructional System Development (ISD) that draws heavily on the work by Tyler and others on the behavioral objectives of instruction.[4] The ISD model consists of the following five steps:

1. Determine job requirements through detailed analysis of tasks performed in an occupational specialty.

2. Determine type of instruction (formal classroom, on-the-job, or other) that best suits the student population and task requirements.

[4]See, for example, R.W. Tyler, Basic Principles of Curriculum and Instruction (Chicago: University of Chicago Press, 1950); and R. W. Tyler, R.M. Gagne, and M. Scriven, Perspectives of Curriculum Evaluation (Chicago: Rand McNally, 1967).

3. Develop objectives that specify the desired behaviors, the conditions under which they are to be demonstrated, and an acceptable standard of performance.

4. Plan and develop instructional methods, media, and equipment.

5. Conduct and evaluate instruction.

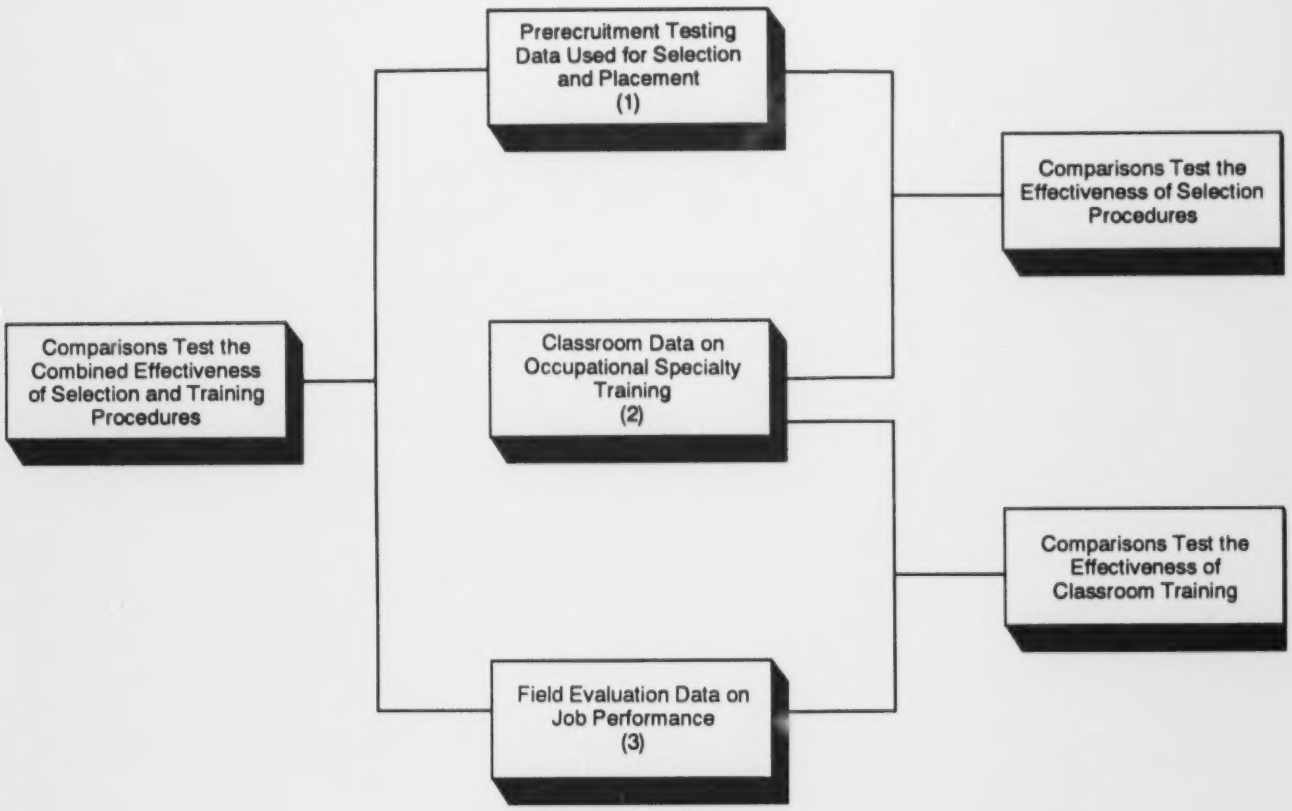
A student's progress through an ISD-developed curriculum is measured by criterion-referenced tests at the end of each block of training. A student passes the course after he or she has performed each task identified as a job requirement at the level of competency defined as acceptable. Continuous monitoring of job requirements is needed to assure that course objectives remain relevant.

Upon successful completion of classroom training in the occupational specialty, the recruit is ready for assignment in the field to carry out the duties requiring the skills acquired during training. Formal training is now complemented by the necessary on-the-job training to permit the recruit to function as part of a unit with a defined mission in a real-world setting.

Objectives, Scope, and Methodology

The purpose of our study is twofold: to profile the aptitudes of the recruits who entered the service from 1981 to 1989, and to evaluate the military service's ability to select successful trainees and to assess their training and work performance. We will examine the three points in a recruit's service career where data critical to performing a thorough evaluation of training must be collected: (1) at entrance to military life, prior to assignment to an occupational specialty; (2) during training, when the recruit's mastery of the specialty's basics is assessed; and (3) after assignment to the field, where what was learned in the classroom must be applied in the work environment. (See figure 1.2.)

Figure 1.2: Data Sources and Comparisons


Prerecruitment Testing Data Used for Selection and Placement (1)
Comparisons Test the Effectiveness of Selection Procedures
Comparisons Test the Combined Effectiveness of Selection and Training Procedures
Classroom Data on Occupational Specialty Training (2)
Comparisons Test the Effectiveness of Classroom Training
Field Evaluation Data on Job Performance (3)

The evaluation model underlying our review assumes the need to interrelate these three points. Comparing the information collected at points 1 and 2 can provide some insight into the ability of the services to predict how well recruits will perform in training on the basis of their scores in qualifying tests. The strength of the relationship between points 2 and 3 is a partial measure of the validity and effectiveness of training. Finally, the relationship between points 1 and 3 is an estimate of the effectiveness of the services' selection and training procedures.

The model is, of course, simplistic and in need of considerable expansion. A fully detailed model would have to consider other influences on performance, such as on-the-job experiences, and would need to be able to determine the location of a problem if relationships between the three

points were weaker than anticipated. Yet, the model, at whatever level of sophistication, would at a minimum require data at these three critical points in a recruit's service career.

We reviewed the information collection practices of each service at the three points identified in the model. For a selected number of occupational specialties—our focus is on training for the more technical occupational specialties—we reviewed the data that have been collected for insights they provide into the service's selection and evaluation procedures, particularly as they affect women and minority groups.

Our study is organized around three evaluation questions, each corresponding to one of the model data points. Each question is addressed in a separate chapter.

1. How has the aptitude of recruits for technologically sophisticated specialties changed since 1980?

DOD tracks recruit aptitude according to four broad mental categories based on the scores on the Armed Forces Qualification Test (AFQT). (See table 1.1.) AFQT is a composite of four of the ten tests from the Armed Services Vocational Aptitude Battery (ASVAB) administered to every potential recruit. We examined some other components of ASVAB in greater detail, particularly those subtests that are used to qualify candidates for high technology occupational specialties.

Table 1.1: How AFQT Test Results Are Categorized

AFQT category	AFQT percentile score	Trainability
I	93–99	Well above average
II	65–92	Above average
IIIA	50–64	Average
IIIB	31–49	Average
IV	10–30	Below average
V[a]	1–9	Well below average

[a]Category V examinees are excluded by law from military service.

2. How useful are the data collected by the services before and during classroom training for selecting individuals for high technology roles and for evaluating the effectiveness of this training?

We examined the measures of recruit performance collected during training and assessed their utility for evaluating training effectiveness,

as well as for providing information on the validity of procedures used to assign recruits to training.

3. How well do the services' selection criteria and training evaluation measures predict success in high technology roles?

We examined the procedures used by each of the services to assess the impact of training on actual job performance. We also related these procedures to the ASVAB scores used to select trainees and to classroom measures of training success, in order to estimate the predictive validity of these measures.

In view of the demographic shifts projected for the labor force over the next decade, we provided separate answers to each of these questions, wherever possible and appropriate, for women and minorities.

We defined high technology roles as those occupational specialties for which the services require a qualifying score in electronics substantially above the mean. For our review, we selected a sample of 13 such courses—five from the Army and four each from the Navy and the Air Force—from which we collected data on individual student performance. Each of these courses is intended to provide a recruit the necessary introductory training to qualify as an apprentice in his specialty.

In the course of our review, we interviewed officials responsible for training evaluation in the Office of the Secretary of Defense and within each of the three services. We visited four service training centers and the facilities maintained by each of the services for research into training and other personnel issues, as well as the Training Performance Data Center in the Office of the Secretary of Defense. Our final data base was compiled from information received from all of these sources, but our primary source for ASVAB and demographic data was the Defense Manpower Data Center. We also received information from the Center for Naval Analyses on technical adjustments to ASVAB validity estimates, and on the ASVAB norm group. This study was conducted in accordance with generally accepted government auditing standards.

Strengths and Limitations of Our Study

Our review of the quality trends among the 2.3 million recruits who entered military service from 1981 to 1989 is more finely grained than the traditional counts of recruits in each of four mental categories routinely reported to the Congress. We report the differences among racial groupings and between male and female recruits, and we examine differential trends among the various areas measured by ASVAB. We assumed the reliability and validity of the widely researched ASVAB and its subtests and made no independent review of these factors. However, we did develop an independent scoring procedure for ASVAB that suggests an alternative, and apparently more valid, approach to assigning recruits to occupational specialties.

The intent of our review of classroom grades and other evaluation measures was to identify the major sources of training evaluation information now in place in the services, and to make use of the objective data we collected to address some concerns about recent trends in recruit quality and the future composition of the recruit pool.

Two important considerations about our sample of students limit any attempt to generalize our findings. First, we deliberately chose occupational specialties for which the services required above average mental qualifications. While the types of classroom measures employed in these courses would most likely be found in other courses with similar requirements, we can say little about the evaluation procedures for less demanding specialties. Second, in part because of the nature of the specialties we chose, our sample contained relatively few members of minority groups and very few women. This fact limited the power of our statistical analysis of these subgroups, and allowed only first-level comparisons (that is, white versus nonwhite; male versus female). Nevertheless, even at this level, we believe we have identified some important differences and gaps in the available data for determining the success of training outcomes. These differences and gaps, together with other findings from our analyses, strongly suggest the need for further, more targeted evaluation of its training efforts by the military.

The Quality of Military Recruits: 1981-89

In 1980, there were 2.4 million more American youths aged 18-21 than there are today. This age group, which now numbers 15 million, will diminish to 13.5 million by the mid-1990's. This 15-year 22-percent decline in the population from which the all-volunteer force draws its new personnel must be a matter of concern to military recruiters. The concern is exacerbated when we consider the technological aptitude of the potential recruit pool: it appears that the graduates of our public schools are becoming less technologically literate when compared to their peers in other developed nations—and this decline is occurring just as our weapons systems are reaching new heights of technological sophistication.

However, by the standards set by DOD, the quality of military recruits in the first half of the 1980's did not decline in proportion to the dwindling numbers in the recruit pool. As we have noted in the previous chapter, DOD reported "the most remarkable turnaround in peacetime history" between 1980 and 1986, with dramatic increases in the proportion of recruits who had graduated from high school and who scored in the top three AFQT categories.

In this chapter, we will address our first evaluation question: How has the aptitude of recruits for technologically sophisticated specialties changed since 1980? Our purpose is threefold: (1) to determine whether the quality gains as defined and reported by the services in the first half of the 1980's are being maintained; (2) to expand the definition of quality to include other measures beyond those traditionally reported (that is, high school graduation and service-defined mental category); and (3) to examine in greater detail two occupational specialties that, by service definition, require higher entry levels of technological sophistication. We will report the trends we found in the scores achieved by recruits from fiscal year 1981 through fiscal year 1989 on some of the various subtests and composites of the Armed Services Vocational Aptitude Battery (ASVAB), the instrument used by all services to both qualify applicants for entry and classify recruits into occupational specialties. We will examine in detail those scores that are used by the services to qualify recruits for more technologically demanding specialties.

Armed Services Vocational Aptitude Battery (ASVAB)

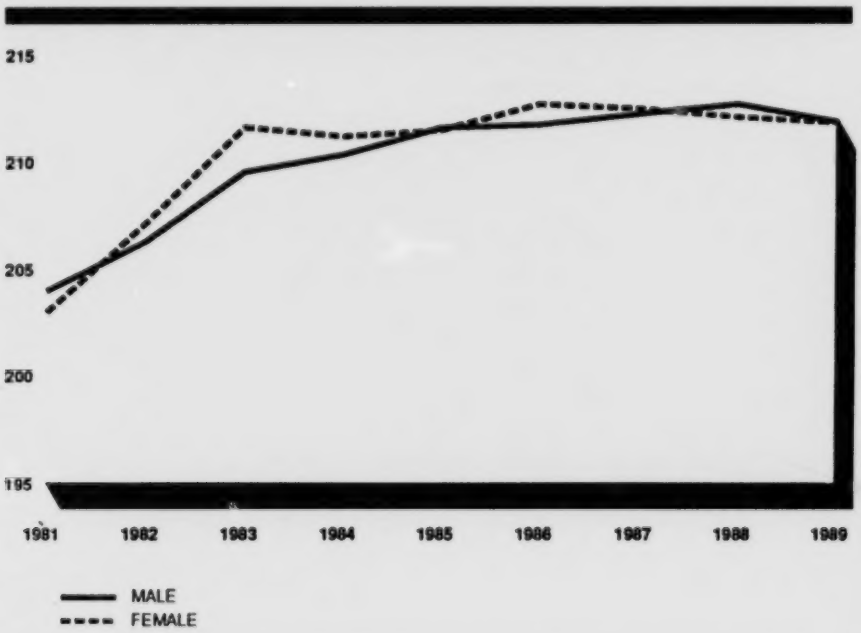
ASVAB is composed of ten subtests measuring abilities considered important for military service. Scores from ASVAB subtests are combined to form composite scores thought to be related to general types of occupational specialties within the armed forces. While different services use different methods to combine subtest scores into composites, all services

use the same component subtests for two composite scores, the Armed Forces Qualification Test (AFQT) and the Electronics Composite. We examined these two in detail to determine how they have changed during the 1980's.

Armed Forces Qualification Test (AFQT)

An AFQT score is currently derived from a recruit's scores on four ASVAB subtests: Word Knowledge, Paragraph Comprehension, Arithmetic Reasoning, and Mathematics Knowledge.[1] AFQT scores are the primary mental criterion for entry into the armed services. Figure 2.1 displays the mean composite AFQT scores for men and women from 1981 through 1989. Actual mean scores for this period may be found in appendix I.

Figure 2.1: Mean AFQT Scores, by Gender: 1981-89



Note: AFQT scores were computed as the sum of standard scores on Arithmetic Reasoning and Mathematics Knowledge, plus the Verbal standard score times two. This is the formula used by DOD as of January 1, 1989.

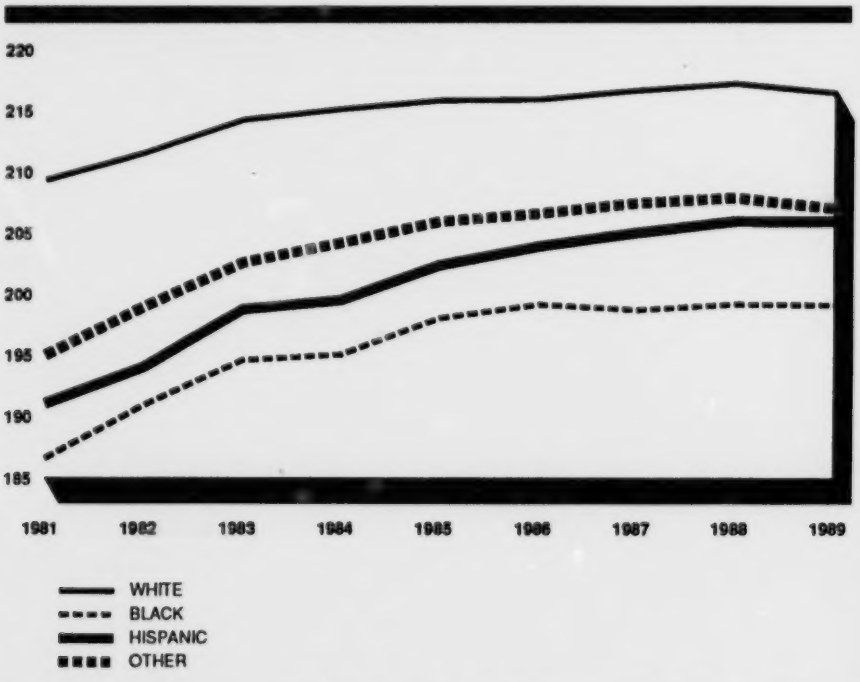
Source: Data are from the Defense Manpower Data Center.

[1]Before 1989, AFQT scores were computed differently. In order to mountain comparability, we computed AFQT scores of all recruits using the 1989 definition and the standard subtest scores provided by the Defense Manpower Data Center.

Overall AFQT scores improved approximately eight points between 1981 and 1989. This improvement occurred among both male and female recruits. However, despite fluctuations over the years, the scores of male recruits began and ended the decade slightly higher than female scores. Male scores continued to increase each year until 1988, although their rate of increase was greatest in the first four years. Female scores improved dramatically from 1981 to 1983 but then flattened out, so that by the end of the decade they were lower than in any year since 1985.

AFQT scores differed more substantially across racial/ethnic groupings than between genders. (See figure 2.2.) White recruits began the decade with scores approximately 21 points higher than minority recruits. By 1989, this difference had shrunk to 15 points. The bulk of the relative gain by minority recruits, however, had occurred by 1985, and any narrowing of this gap since then has been slight.

Figure 2.2: Mean AFQT Scores, by Race/ Ethnicity: 1981-89

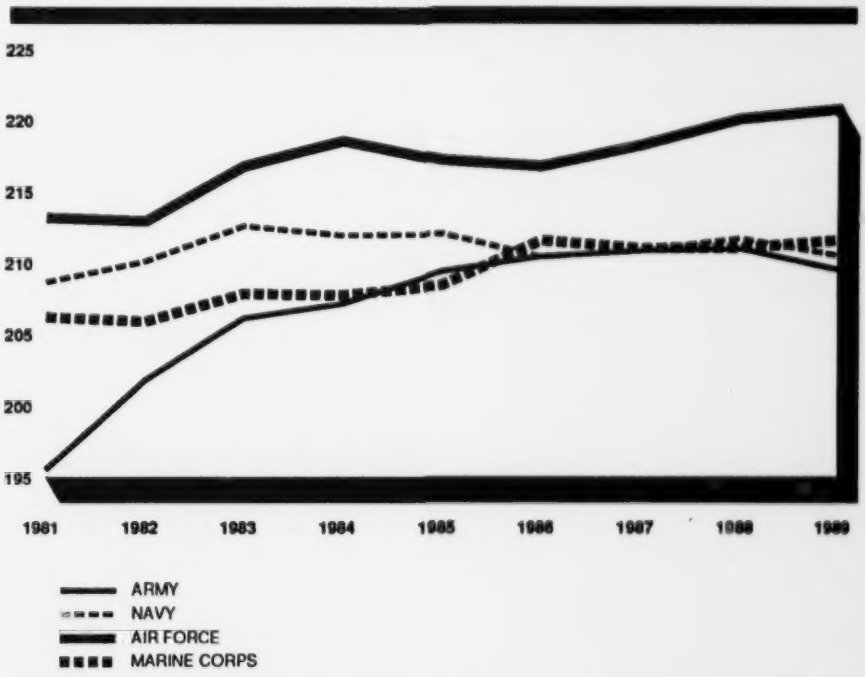


Note: AFQT scores were computed as the sum of standard scores on Arithmetic Reasoning and Mathematics Knowledge, plus the Verbal standard score times two. This is the formula used by DOD as of January 1, 1989.

Source: Data are from the Defense Manpower Data Center.

Mean AFQT scores in all services were significantly higher in 1989 than in 1981. (See figure 2.3.) Army recruits showed the greatest gain. Average Army scores were substantially lower than those of other services at the beginning of the decade, but by 1986 they had increased to approximately the same level as scores achieved by Navy and Marine recruits. Navy scores peaked in 1983 and have declined somewhat slowly and erratically since then to a level less than 2 points higher than they were at the beginning of the decade. Air Force AFQT scores have consistently averaged higher than the other services' and have not displayed their tendency to plateau at mid-decade levels.

Figure 2.3: Mean AFQT Scores, by Service: 1981-89



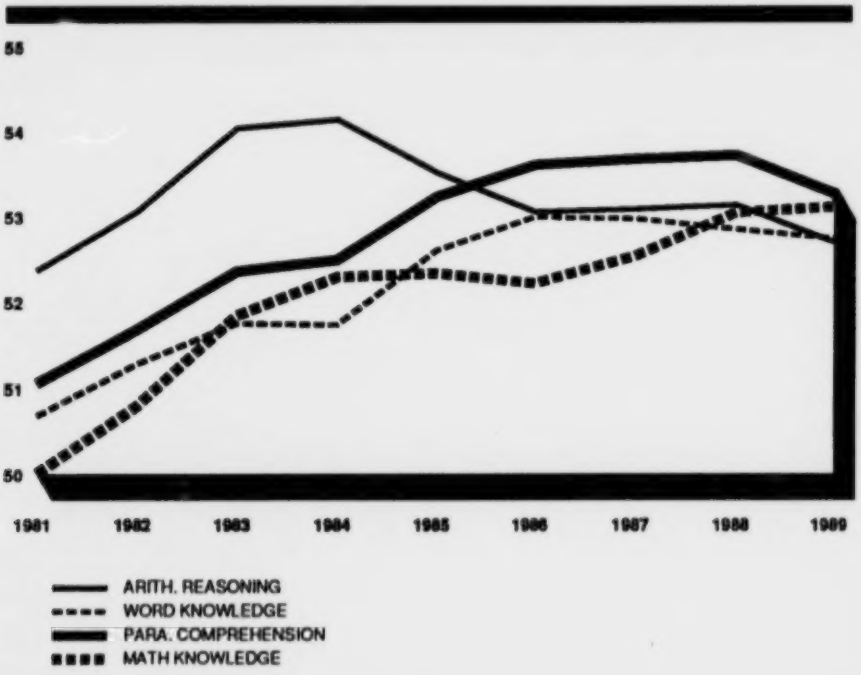
Note: AFQT scores were computed as the sum of standard scores on Arithmetic Reasoning and Mathematics Knowledge, plus the Verbal standard score times two. This is the formula used by DOD as of January 1, 1989.

Source: Data are from the Defense Manpower Data Center.

Figure 2.4 displays the service-wide mean scores on each of the four component subtests that make up AFQT. For two of the subtests, Word Knowledge and Paragraph Comprehension, the pattern is quite similar, with the sharpest gains occurring by 1985, and little change thereafter.

Scores in Mathematics Knowledge and Arithmetic Reasoning increased substantially between 1981 and 1984. Arithmetic Reasoning scores declined after that point, but scores in Mathematics Knowledge have continued to rise and were the only subtest scores to increase from fiscal year 1988 to fiscal year 1989.

Figure 2.4: Mean AFQT Subtest Scores, 1981-89

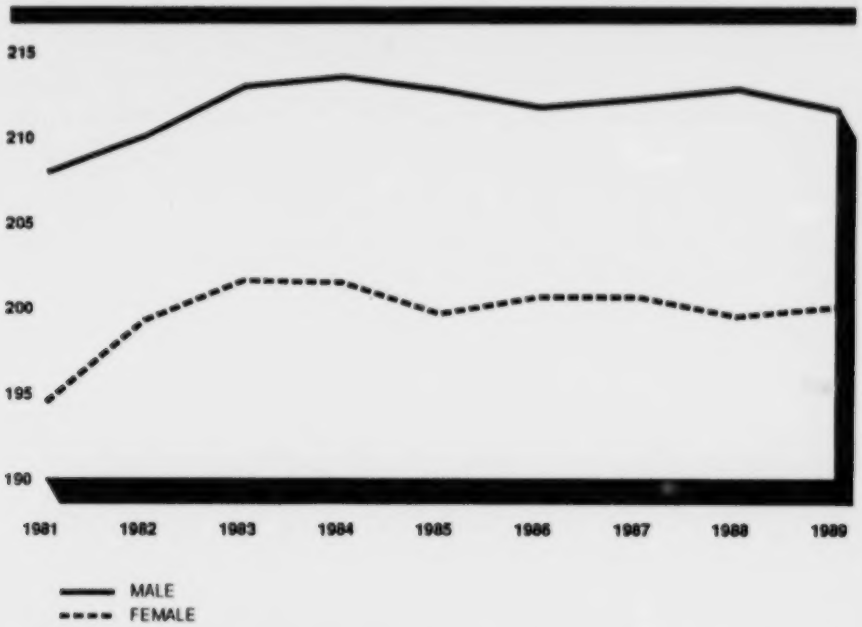


Source: Data are from the Defense Manpower Data Center.

Electronics Composite Scores

The Electronics Composite score is defined by each service as the sum of four subtest scores: Arithmetic Reasoning, Mathematics Knowledge, Electronics Information, and General Science. Figure 2.5 displays the mean Electronics Composite score for men and women from 1981 through 1989. Figure 2.6 presents the same information by racial/ethnic grouping.

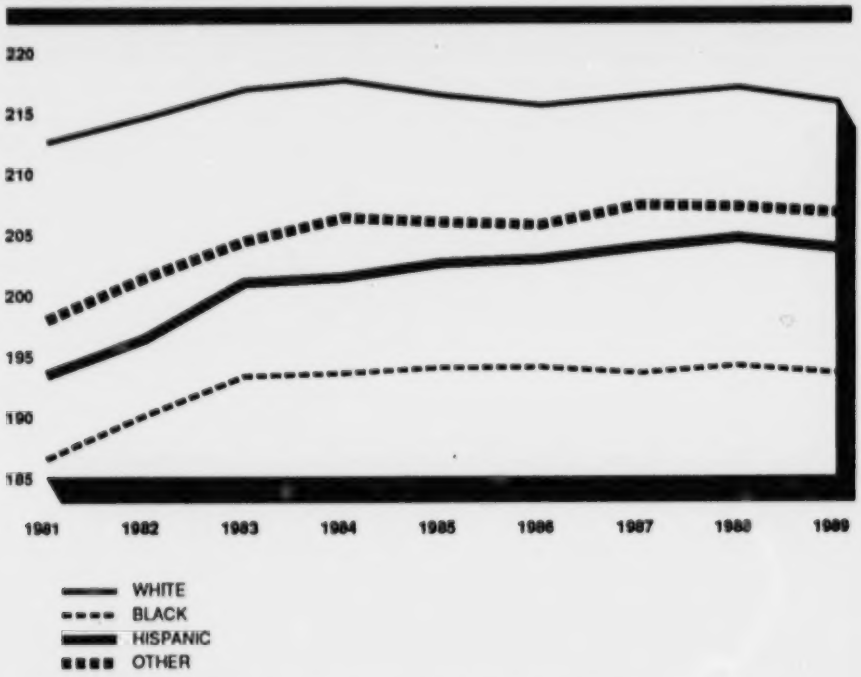
Figure 2.5: Mean Electronics Composite Scores, by Gender: 1981-89



Note: Electronics Composite scores were computed as the sum of standard scores on Arithmetic Reasoning, Mathematics Knowledge, Electronics Information, and General Science.

Source: Data are from the Defense Manpower Data Center.

Figure 2.6: Mean Electronics Composite Scores, by Race/Ethnicity: 1981-89



Note: Electronics Composite scores were computed as the sum of standard scores on Arithmetic Reasoning, Mathematics Knowledge, Electronics Information, and General Science.

Source: Data are from the Defense Manpower Data Center.

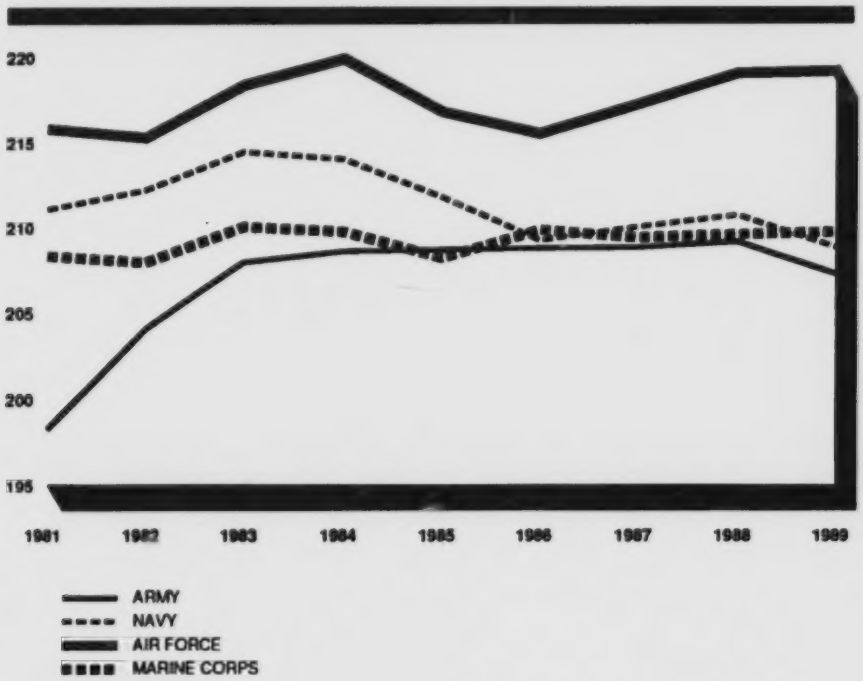
Electronics Composite mean scores rose approximately 3-1/2 points between 1981 and 1989. They peaked in 1984 and experienced a gradual decline thereafter. Female recruits scored approximately 11 points lower than male recruits during this period.

Because of the overlap between the Electronics Composite and AFQT, the racial differences are similar. In 1981, white recruits scored approximately 24 points higher than minorities on this composite. By 1989, the gap had narrowed to approximately 19 points, but most of these gains by minorities were attained in the earlier part of the decade. By 1989, the scores of all racial groups were declining.

The interservice pattern of Electronics Composite scores is again similar to the AFQT patterns discussed previously. (See figure 2.7.) Army scores progressed from an average of ten points lower than the next closest service in 1981 to being essentially the same as Navy and Marine scores

by 1986. Mean scores for these three services changed very little from 1985 to 1988, but Army and Navy scores declined significantly in 1989. Air Force scores have remained higher than other services' but have fluctuated irregularly since 1984.

Figure 2.7: Mean Electronics Composite Scores, by Service: 1981-89

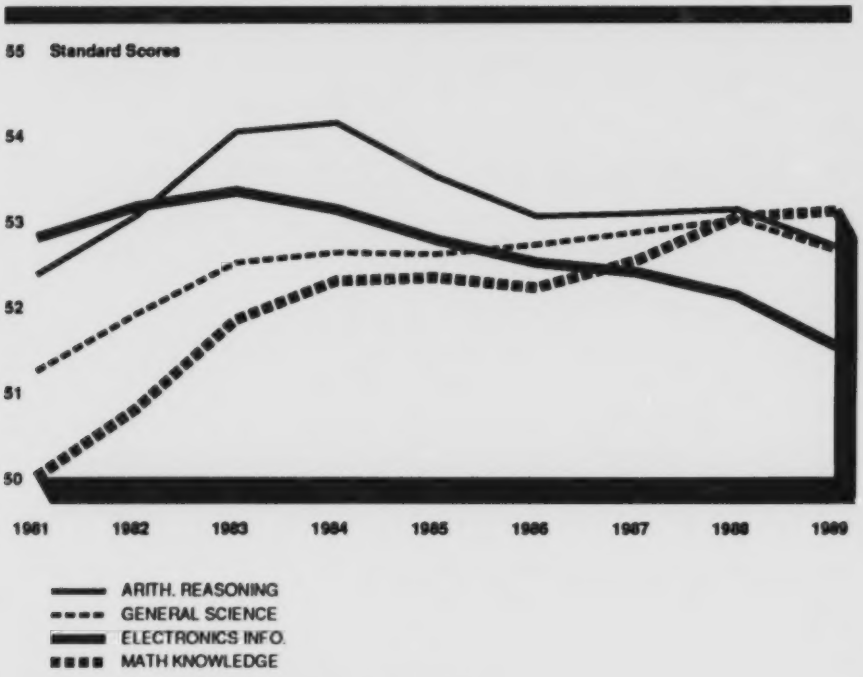


Note: Electronics Composite scores were computed as the sum of standard scores on Arithmetic Reasoning, Mathematics Knowledge, Electronics Information, and General Science.

Source: Data are from the Defense Manpower Data Center.

The trends during this period were not the same for all the subtests that comprise the Electronics Composite score. (See figure 2.8.) Scores in General Science and Mathematics Knowledge increased steadily over these years. Scores in Arithmetic Reasoning increased from 1981 to 1983 but by 1986 had declined again and have since remained relatively constant. In 1981, recruits scored higher in Electronics Information than in the other component subtests, but by 1988 the scores were lower than for other subtests and lower even than they had been at the beginning of the decade. In 1989, they declined further.

Figure 2.8: Mean Electronics Composite Subtest Scores, 1981-89



Source: Data are from the Defense Manpower Data Center.

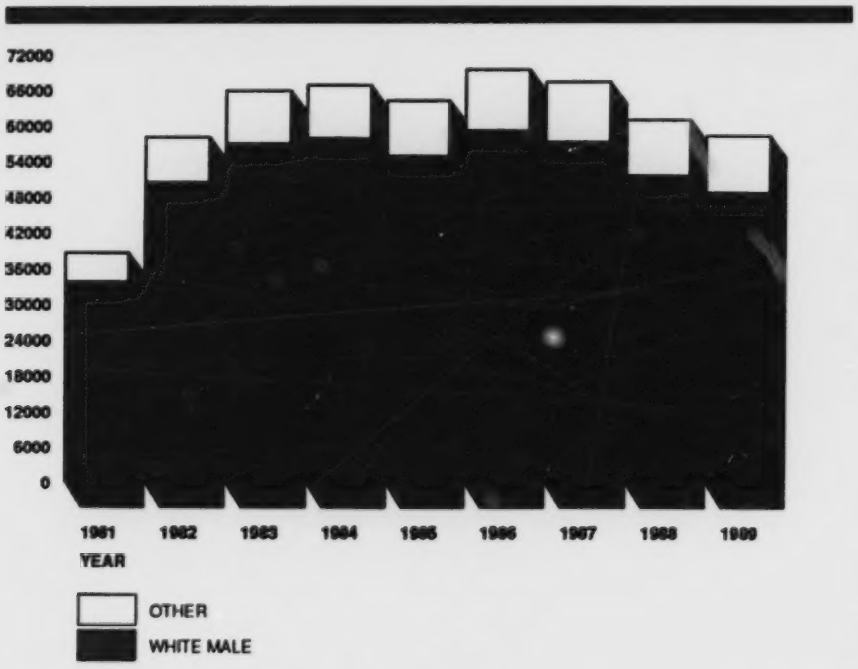
Number of Recruits Qualified for High Technology Specialties

An alternative method for examining trends in recruit qualifications is to enumerate the number of recruits whose ASVAB scores meet the minimum standards required for entry into certain occupational specialties. Each service defines "cutting scores" for classifying recruits—that is, a minimum score on one or more ASVAB composites is required for entry into training for each specialty.[2] This score can be adjusted to control flow into specialties as needed. We chose two of the more demanding specialties, both of them in the Air Force, and computed the number of recruits into each service from 1981 to 1989 whose ASVAB scores would have qualified them for technical training in these specialties. We chose these specialties as examples of high technology military occupations because they share cutting scores with a number of other technologically oriented specialties. Our purpose was not to imply either a surplus or deficit of requisite manpower.

[2]Other qualifications may also apply—for example, possession of a valid driver's license, special physical qualifications, or the ability to obtain appropriate levels of security clearance.

Figure 2.9 depicts the number of recruits during the period in question who would have qualified for training as control and warning radar specialists in the Air Force on the basis of their ASVAB scores.[3] In 1981, approximately 38,000 recruits qualified for this specialty. By 1986, the number of recruits qualifying had risen to more than 69,000, but since then the number has declined to just under 58,000. In 1981, 87 percent of the recruits qualifying for training as control and warning radar specialists were white males, although only about two thirds of 1981 recruits were white males. These proportions had not changed substantially by 1989, when white males comprised 84 percent of qualified recruits but only 61 percent of the general recruit population.

Figure 2.9: Number of Recruits Qualifying for Training as Control and Warning Radar Specialists, 1981-89



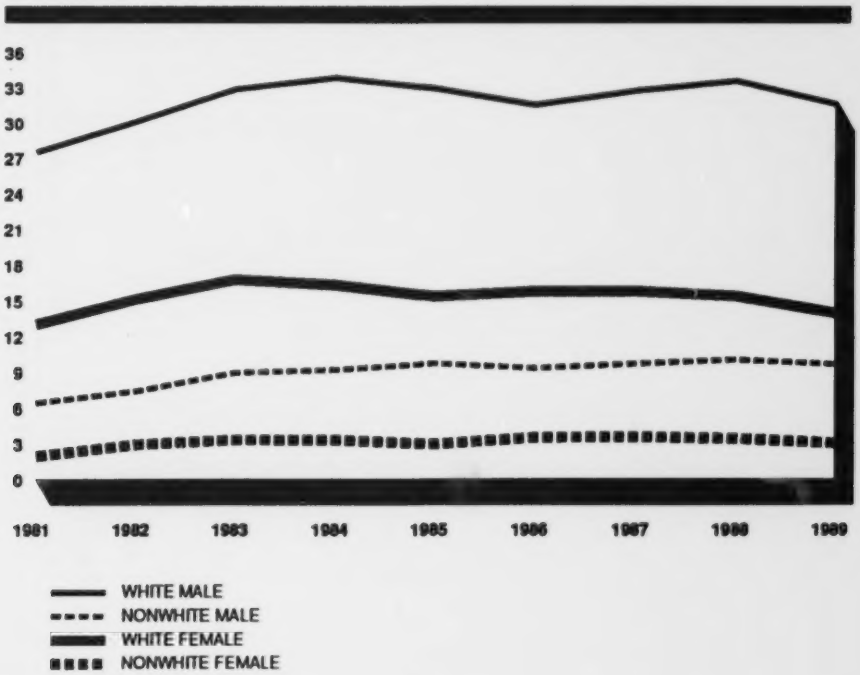
Source: Data are from the Defense Manpower Data Center.

Because the total manpower quotas for the services have varied over this period, we also computed the percent of all recruits within the

[3]We used the cutting score that was current for Air Force recruits in May 1989—an Electronics Composite score of 230.

gender and racial/ethnic groups who qualified for this specialty. The results are displayed in figure 2.10.

Figure 2.10: Percent of Recruits Qualifying for Training as Control and Warning Radar Specialists, 1981-89

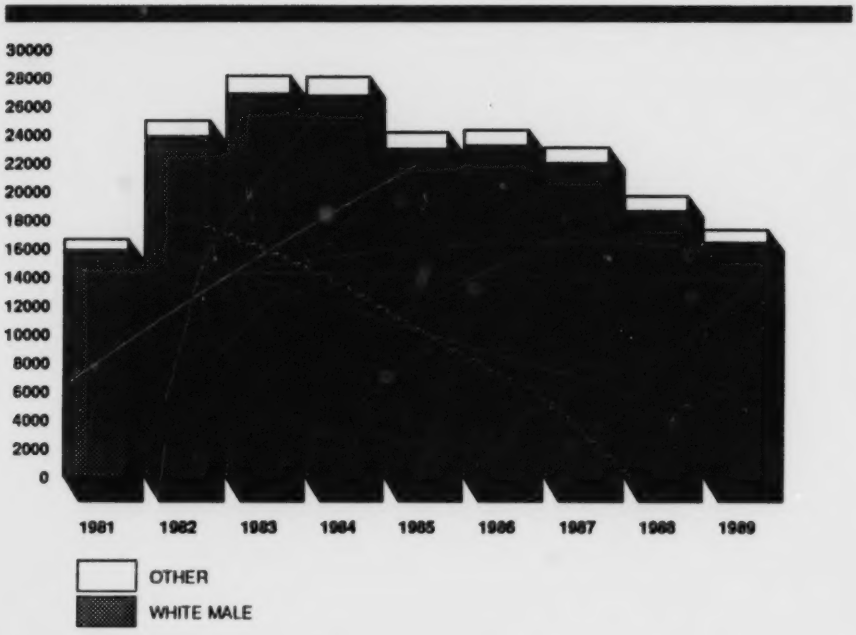


Source: Data are from the Defense Manpower Data Center.

While nearly a third of white males who entered the services during this period qualified on the basis of their Electronics Composite scores for this occupational specialty, fewer than 15 percent of white females qualified. Fewer than 10 percent of minority males and approximately 3 percent of minority females qualified.

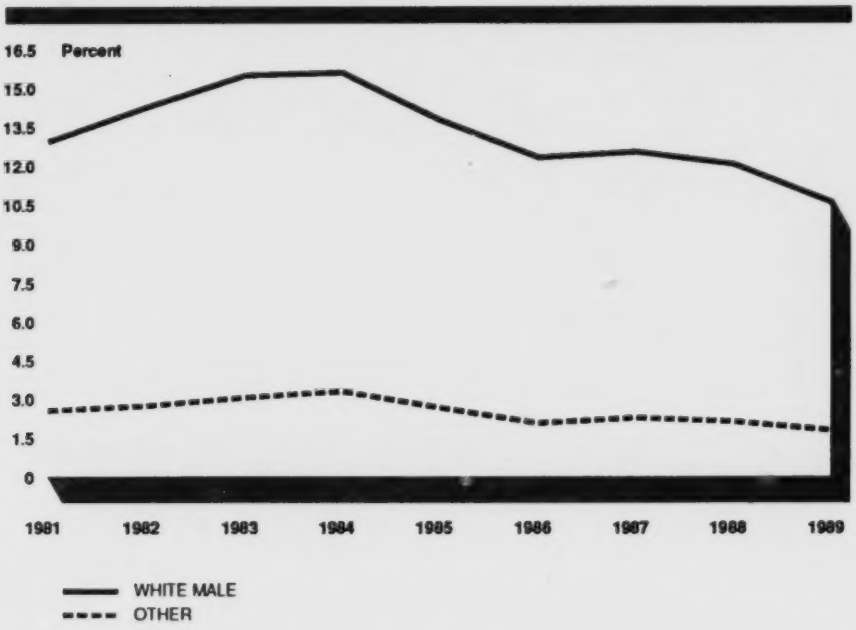
The demographic differences are even more sharply defined when the occupational specialty of Systems Repair Technician is examined. (See figures 2.11 and 2.12.)

Figure 2.11: Number of Recruits Qualifying for Training as Systems Repair Technicians, 1981-89



Source: Data are from the Defense Manpower Data Center.

Figure 2.12: Percent of Recruits Qualifying for Training as Systems Repair Technicians, 1981-89



Source: Data are from the Defense Manpower Data Center.

In 1981, 16,563 recruits met the demanding qualifications for training in this field.[4] The number of qualified recruits increased sharply by 1983, but by the end of the decade it had dropped to within 700 of its 1981 level. The vast majority of these were white males, of whom approximately 11 percent qualified. Fewer than 2 percent of our other demographic groups met the qualifications.

Summary and Conclusions

As we approach the twenty-first century, the sophistication of our weapons systems can be expected to impose greater demands on the technological competence of the individual members of the armed forces. In addition, the youth pool from which the services will draw their recruits will become increasingly female and minority. And although we cannot foresee how reduced political tensions may ease the demands on this pool, our examination of recruit quality trends during the 1980's is not reassuring concerning the military's ability to meet these challenges.

[4]This specialty requires an ASVAB Electronics Composite score of 235 and a mechanical score of 247, requirements that rank it among the most challenging fields in all of the services.

AFQT scores and, to a lesser extent, Electronics Composite scores are higher now than they were in 1981, yet both have begun to decline. The Electronics Information subtest scores are lower than they were in 1981, and General Science scores have dropped to near their 1981 level. Thus, fewer recruits are qualifying for the more demanding technical occupational specialties.

Women and minorities have traditionally scored lower in these areas. While the gap between white males and other recruits narrowed somewhat in the early 1980's, since mid-decade the race and gender differences have remained fairly constant. As we discussed in the previous chapter, women and minorities will form the bulk of the new-entry labor pool by the year 2000, and therefore providing well-trained personnel for a technologically sophisticated military can be expected to become increasingly difficult. The burden on training will increase, and with it will come the need to monitor the effectiveness of this training as recruit demographics shift.

In the following chapters, we will address the services' current ability to measure the effectiveness of their training in technologically demanding areas. We will also examine the differences among gender and racial/ethnic groupings, and the ability of the AFQT and Electronics Composite scores to predict success in technical military specialties.

Classroom Measures of Training Effectiveness

In this chapter, we address our second evaluation question: How useful are the data collected by the services before and during classroom training for selecting individuals for high technology roles and for evaluating the effectiveness of this training? Although we reviewed a broad spectrum of evaluation-related materials and activities performed by the services at the classroom level, we concentrated on the course grades assigned at the end of training and, in some cases, at intermediate stages during the training process. Our intention was to define the extent to which appropriate data were available to the services and to external reviewers from which some judgments could be made about training effectiveness. We did not attempt to perform an evaluation of individual curricula, training sites, or instructors.

Our primary criterion for selecting courses for review was that the qualifying score for course entry, as established by the service, was relatively high. In addition, we considered annual trainee throughput and the recent stability of the course curriculum. Nearly all the courses which met our criteria were in the electronics area, and most involved the use, maintenance, and repair of electronic equipment, particularly radar or sonar. We collected the course grades associated with advanced individual training for 13 occupational specialties, four each in the Navy and Air Force, and five in the Army. Some of the data were collected at the training site, and some from centrally computerized records.

Because of large differences between the services in annual throughput of trainees in these courses, the size of our sample varied widely across services. This variation was increased by problems we encountered concerning the usefulness of certain data provided by the Army (see the following section), as well as by our decision to supplement our already sizable Navy data base with relevant data previously collected by the Navy for research purposes. Our final sample consisted of more than 6,000 sailors, nearly 1,000 Air Force personnel, and fewer than 300 soldiers. In this chapter, we present the results of our analysis separately for each service.

We examined the course data for their apparent reliability—that is, for their apparent ability to discriminate meaningfully between performances of trainees—as well as for differences in training outcomes among the demographic groupings discussed in the previous chapter. We also examined the relationship between training outcomes and individual abilities, as measured by ASVAB, in order to estimate the power of the selection criteria to predict performance in training.

Army

The Army specialties for which we collected data are listed in table 3.1.

Table 3.1: Army Occupational Specialties Reviewed

Specialty	Title	Location	Electronics Composite qualifying score[a]
24J	Hawk pulse radar repairer	Redstone Arsenal, Ala.	217
27N	Forward area alerting radar repairer	Redstone Arsenal, Ala.	217
29V	Strategic microwave systems repairer	Fort Gordon, Ga.	217
36L	Transportable automatic systems operator	Fort Gordon, Ga.	217
39B	Automatic test equipment operator	Fort Gordon, Ga.	217

[a]Sum of subtest standard scores

We found that the course grades for these five specialties were not equally reliable indicators of performance during training. Whereas for the two classes at Redstone Arsenal final grades were a simple arithmetic average of intermediate measures of performance, at Fort Gordon we were unable to find a consistent relationship between individual milestone measures and final grades, nor were we able to locate anyone at Fort Gordon who could suggest one. We concluded that the grades recorded for two of these courses (36L and 39B) could not be used to discriminate reliably between the performances of individual trainees. We found inconsistencies in scoring procedures between different classes and even within the same class. Finally, we discovered that the Fort Gordon grades (unlike those at Redstone) were based partially on measures of physical conditioning that appeared to be unrelated to job performance.

For a third training course at Ford Gordon (29V), however, we were able to generate what we judged to be reasonable measures of performance for some classes. For these classes, we developed an algorithm to produce scores based only on those nonconstant measures that were related to general or applied electronics training.[1]

[1]External corroboration of the preferability of this improvised scoring procedure was provided by our later analysis of the relationship between grades and ASVAB. The correlation between original 29V grades and the Electronics Composite was negative and nonsignificant. The revised grades were positively (.50) and significantly correlated ($p < .01$) with this ASVAB score.

Our final sample was therefore composed of U.S. Army trainees from those 24J and 27N classes conducted in fiscal years 1985 through 1988 whose records were available at the time of our visit, and approximately one third of the 29V trainees from the same period. Table 3.2 presents the mean scores of this sample on AFQT, the Electronics Composite of ASVAB, and course grades.[2]

Table 3.2: Mean Scores on Predictor and Criterion Variables, Army

	AFQT		Electronics Composite		Grade	
Category	**Number**	**Mean[a]**	**Number**	**Mean[a]**	**Number**	**Mean**
Male	280	232.15	280	238.46	232	89.23
Female	23	232.87	23	230.13	23	86.08
White	255	234.00	255	240.00	160	90.19
Nonwhite	48	222.67	48	226.29	95	86.86
Total	**303**	**232.20**	**303**	**237.83**	**255**	**88.95**

[a]Sum of subtest standard scores

Male trainees in these courses scored significantly higher than did females, and white trainees performed better than minority students. These performance differences correspond to group-level differences in both AFQT and Electronics Composite scores for racial/ethnic groupings.

The group means presented in table 3.2 also suggest that AFQT and Electronics Composite scores do not equally predict success in training, at least for females. While female trainees entered training with Electronics Composite scores significantly lower than those of males, the AFQT scores of female and male trainees were equivalent. In other words, it would appear that Electronics Composite scores are a better indication of future performance in these occupational specialties than are AFQT scores. This is consistent with ASVAB's role in the military accession process: potential recruits are admitted to service on the basis of AFQT scores, and then are assigned to occupational specialties for which they qualify on the basis of their scores on other ASVAB composites.

We tested this hypothesis more directly by examining the correlations between course grades and three ASVAB scores: AFQT, Electronics Composite, and a "factor score." This last measure is the weighted sum of all ten ASVAB subtests. We derived this last score by principal component analysis of ASVAB subtest scores. The results of our correlation analysis are displayed in table 3.3.

[2]See appendix II for similar statistics on the course level.

Table 3.3: Intercorrelation of Study Variables, Army[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.81[g]	0.84[g]	0.29[g]	0.41[g]
Electronics Composite	303	1.00	0.89[g]	0.43[g]	0.59[g]
Factor	303	303	1.00	0.42[g]	
Grade	189	189	189	1.00	
Male					
AFQT	1.00	0.83[g]	0.85[g]	0.31[g]	0.43[g]
Electronics Composite	280	1.00	0.89[g]	0.42[g]	0.58[g]
Factor	280	280	1.00	0.41[g]	
Grade	171	171	171	1.00	
Female					
AFQT	1.00	0.82[g]	0.87[g]	0.42	0.53[g]
Electronics Composite	23	1.00	0.89	0.35	0.51[g]
Factor	23	23	1.00	0.35	
Grade	18	18	18	1.00	
White					
AFQT	1.00	0.80[g]	0.82[g]	0.24[g]	0.38[g]
Electronics Composite	255	1.00	0.87[g]	0.40[g]	0.60[g]
Factor	255	255	1.00	0.40[g]	
Grade	154	154	154	1.00	
Nonwhite					
AFQT	1.00	0.78[g]	0.85[g]	0.19	0.22
Electronics Composite	48	1.00	0.89[g]	0.30	0.40
Factor	48	48	1.00	0.26	
Grade	35	35	35	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal.

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

For our whole Army sample, the variation within Electronics Composite scores explains approximately 18 percent of the variation within course

grades, more than factor scores and substantially more than AFQT.[3] In most cases, Electronics Composite scores are somewhat better predictors of grades than are AFQT scores, whether a simple correlation coefficient or a coefficient adjusted for range restriction is used as a criterion.[4] This is not true, however, for female soldiers, for whom AFQT predicts classroom performance better than the Electronics Composite does. In most cases, ASVAB factor scores provide stronger predictions than either AFQT or the Electronics Composite. Our ability to predict course grades from any of the three ASVAB scores is weakest for minority soldiers as a group.

Our analysis of nonwhite and female soldiers is unfortunately based on a relatively small sample. Nevertheless, it suggests that AFQT or some other general score from ASVAB may provide a better predictor of success for women recruits in electronics-related training than does the Electronics Composite score. It also indicates that we need better predictors than we currently have for minority students.

Navy

We examined four Navy training courses, two each from the Antisubmarine Warfare School in San Diego and the Naval Air Station in Millington, Tennessee. They are listed in table 3.4.

[3]A correlation coefficient is the square root of common variance. In this case, the Electronics Composite score from ASVAB shares 18.5 percent ($.43^2$) of variance with grades, or, after adjustment, 35 percent ($.59^2$).

[4]The adjustment for restriction in range is common among psychometricians and appears in all DOD reports that we reviewed. Since correlations are simply measures of the extent to which two measures vary in common, any restriction to the variation of one of the measures results in an underestimate of their common variation. This restriction occurs when the sample includes only one end of a spectrum of scores, as is the case for any measure used for selection purposes. Our sample includes only those whose AFQT scores were sufficiently high to permit acceptance into military service. The adjusted correlation coefficient represents the hypothetical relationship between the ASVAB measure and course grades if this range restriction did not exist for our sample.

Table 3.4: Occupational Specialties Reviewed, Navy

Specialty	Title	Location	Electronics Composite qualifying score[a]
STG	Sonar technician, antisubmarine warfare, surface	San Diego, Calif.	218
STS	Sonar technician, antisubmarine warfare, subsurface	San Diego, Calif.	218
AQ	Aviation fire control technician	Millington, Tenn.	218
AX	Aviation antisubmarine warfare technician	Millington, Tenn.	218

[a]Sum of subtest standard scores

We were able to achieve a much larger sample size (6,156) for these courses than was the case for our Army courses (303) because of their larger annual throughput, and because the Naval Personnel Research and Development Center provided us with relevant data that they had collected on STS and STG specialties for fiscal years 1986 and 1987. These data supplemented the fiscal year 1988 and fiscal year 1989 data that we collected at the San Diego base. Millington provided us with training data for 1987 and 1988. Table 3.5 presents the mean scores on the two ASVAB composites and course grades for the entire Navy sample. Statistics on individual courses are presented in appendix II.

Table 3.5: Mean Scores on Predictor and Criterion Variables, Navy

	AFQT		Electronics Composite		Grade	
Category	**Number**	**Mean[a]**	**Number**	**Mean[a]**	**Number**	**Mean**
Male	6,080	229.60	6,080	235.33	5,882	89.11
Female	76	235.59	76	230.66	71	90.70
White	5,355	230.49	5,355	236.25	5,179	89.21
Nonwhite	801	224.18	801	228.75	1,159	89.58
Total	**6,156**	**229.67**	**6,156**	**235.28**	**6,443**	**89.30**

[a]Sum of subtest standard scores

Male recruits entered training with significantly lower AFQT scores and significantly higher Electronics Composite scores than those for females. Final grades for males were slightly, but significantly, lower than those for their female classmates. These results suggest that, at least for females, a substantial advantage in AFQT can overcome a disadvantage in the Electronics Composite. In addition, minority students began

training with substantially lower scores than nonminorities on both AFQT and the Electronics Composite. The final grades of the two groups were not significantly different.

The results of our correlation analysis appear in table 3.6. They suggest that AFQT may be more important for training success than the Electronics Composite. For most Navy groupings, AFQT scores are better predictors of classroom performance than are Electronics Composite scores. When adjusted, they explain from 12 to 38 percent of the variation in course grades. Once again, the Electronics Composite is the weakest of the three predictors for female sailors, and the more general factor score is the strongest. The ability of any of the three ASVAB scores to predict training success is weakest for minorities.

Table 3.6: Intercorrelation of Study Variables, Navy[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.79[g]	0.80[g]	0.30[g]	0.46[g]
Electronics Composite	6,156	1.00	0.85[g]	0.27[g]	0.46[g]
Factor	6,156	6,156	1.00	0.28[g]	
Grade	5,939	5,939	5,939	1.00	
Male					
AFQT	1.00	0.79[g]	0.81[g]	0.30[g]	0.46[g]
Electronic Composite	6,080	1.00	0.85[g]	0.27[g]	0.46[g]
Factor	6,080	6,080	1.00	0.27[g]	
Grade	5,868	5,868	5,868	1.00	
Female					
AFQT	1.00	0.74[g]	0.81[g]	0.39[g]	0.62[g]
Electronics Composite	76	1.00	0.82[g]	0.32[g]	0.55[g]
Factor	76	76	1.00	0.39[g]	
Grade	71	71	71	1.00	
White					
AFQT	1.00	0.79[g]	0.81[g]	0.30[g]	0.47[g]
Electronics Composite	5,355	1.00	0.85[g]	0.29[g]	0.50[g]
Factor	5,355	5,355	1.00	0.30[g]	
Grade	5,165	5,165	5,165	1.00	
Nonwhite					
AFQT	1.00	0.74[g]	0.77[g]	0.22[g]	0.34[g]
Electronics Composite	801	1.00	0.81[g]	0.14[g]	0.25[g]
Factor	801	801	1.00	0.11[g]	
Grade	774	774	774	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal.

[b]AFQT = sum of subtest standard scores.

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite.

[d]Factor = score from first factor from principal component analysis.

[e]Grade = final course grade.

[f]Adjusted = correlation adjusted for restriction of range.

[g]$p < .05$

Air Force

The four Air Force training courses we reviewed are listed in table 3.7. Our sample size from these courses totaled 922. Statistics for individual courses are provided in appendix II. (We received both training and

demographic data on all of these courses from the Air Force Human Resources Laboratory.)

Table 3.7: Occupational Specialties Reviewed, Air Force

Specialty	Title	Location	Electronics Composite qualifying score[a]
30332	Aircraft control and warning radar specialist	Keesler AFB, Miss.	230
30333	Automatic tracking radar specialist	Keesler AFB, Miss.	225
45530A	Photo-sensors maintenance specialist, tactical reconnaissance sensors	Lowry AFB, Colo.	225
45530B	Photo-sensors maintenance specialist, reconnaissance electro-optical sensors	Lowry AFB, Colo.	225

[a]Sum of subtest standard scores

Trainees' ASVAB scores and course grades are displayed in table 3.8. As would be expected, ASVAB scores for Air Force students are significantly higher than those for the other services we reviewed. In addition, we found a higher proportion of female trainees in the Air Force courses than in the Army and Navy courses we reviewed.

Table 3.8: Mean Scores on Predictor and Criterion Variables, Air Force

	AFQT		Electronics Composite		Grade	
Category	Number	Mean[a]	Number	Mean[a]	Number	Mean
Male	824	235.45	824	241.94	854	91.31
Female	98	237.73	98	235.88	100	89.91
White	825	236.22	825	241.95	855	91.21
Nonwhite	97	231.19	97	235.73	99	90.76
Total	**922**	**235.69**	**922**	**241.30**	**954**	**91.16**

[a]Sum of subtest standard scores

Male Air Force recruits entered training with substantially higher Electronics Composite scores and slightly, but significantly, lower AFQT scores than did female recruits. Despite the slight female AFQT advantage, male recruits ended training with higher course grades than those earned by female recruits. In addition, although white students began training with substantially higher ASVAB scores, their final grades were not significantly different from those of their nonwhite classmates.

As table 3.9 demonstrates, the correlations between ASVAB and Air Force training grades followed much the same pattern as did the Navy's. When correlations are adjusted, the traditional ASVAB composite scores explain from 6 to 36 percent of classroom performance. Factor scores are as good as, or better than, composites as predictors. For female students, AFQT scores outpredict Electronics Composite scores. Once again, it is most difficult to predict course grades for minority students, although factor scores explained 10 percent of their classroom performance.

Table 3.9: Intercorrelation of Study Variables, Air Force[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.71[g]	0.75[g]	0.29[g]	0.44[g]
Electronics Composite	922	1.00	0.84[g]	0.33[g]	0.54[g]
Factor	922	922	1.00	0.35[g]	
Grade	922	922	922	1.00	
Male					
AFQT	1.00	0.74[g]	0.77[g]	0.30[g]	0.44[g]
Electronics Composite	824	1.00	0.84[g]	0.33[g]	0.54[g]
Factor	824	824	1.00	0.34[g]	
Grade	824	824	824	1.00	
Female					
AFQT	1.00	0.68[g]	0.77[g]	0.35[g]	0.54[g]
Electronics Composite	98	1.00	0.77[g]	0.26[g]	0.50[g]
Factor	98	98	1.00	0.28[g]	
Grade	98	98	98	1.00	
White					
AFQT	1.00	0.72[g]	0.75[g]	0.31[g]	0.47[g]
Electronics Composite	825	1.00	0.83[g]	0.35[g]	0.58[g]
Factor	825	825	1.00	0.35[g]	
Grade	825	825	825	1.00	
Nonwhite					
AFQT	1.00	0.65[g]	0.68[g]	0.19	0.24[g]
Electronics Composite	97	1.00	0.82[g]	0.23[g]	0.33[g]
Factor	97	97	1.00	0.31[g]	
Grade	97	97	97	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal.

[b]AFQT = sum of subtest standard scores.

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite.

[d]Factor = score from first factor from principal component analysis.

[e]Grade = final course grade.

[f]Adjusted = correlation adjusted for restriction of range.

[g]$p < .05$

Summary and Conclusions

Our review of advanced individual training courses—designed to prepare recruits in three services to serve in certain "high technology" roles—identified some problems with the utility of data maintained by the Army on classroom performance in certain specialties. It would not

be appropriate to make interservice comparisons on the basis of this finding, however, since much of the Navy training information and all of the data we received from the Air Force were specially prepared for research purposes. We cannot therefore make firm judgments about the immediate availability of psychometrically suitable measures from these two services.

The psychometric deficiencies we found at Fort Gordon appeared to result from a number of different factors, including questionable data entry procedures and software. They are also a function of the pass/fail nature of the criteria used to evaluate student progress. We cannot assess the extent to which performance on individual training tasks is susceptible to more sophisticated measures than "go/no-go," but we would suggest that subject matter experts attempt to develop more finely tuned, objective, and reliable measures of performance.

Our review also raised certain questions about differential success in training for males and females, and for whites and minorities, and about the differential predictive validity of ASVAB for these subgroups. Our analysis of gender- and race-related differences in mean ASVAB scores and course grades in the Army suggested that the Electronics Composite was an efficient simple predictor of training success. Women and minorities entered training with significantly lower Electronics Composite scores and received significantly lower course grades.

Our findings from the Navy and Air Force samples, however, suggest that a more complex relationship exists between ASVAB and course grades. For these services, gender- and race-related differences in course grades were small or nonexistent, despite significant differences in Electronics Composite scores. The Navy and Air Force samples also differed from the Army sample in three other respects: (1) Electronics course grade differences, though significant, were much smaller in the Navy and Air Force than in the Army; (2) unlike women soldiers, Navy and Air Force women had significantly higher AFQT scores than their male classmates; and (3) the AFQT disadvantage for minorities in the Navy and Air Force was only half of that in the Army. These findings suggest that an advantage in the more general aptitude measured by AFQT (or by an even more general measure such as a factor score) can compensate for a deficit in the Electronics Composite when the deficit is not too great. In other words, success in training may be related as much to general ability as to performance on the Electronics Composite.

This interpretation is consistent with the results of our correlation analyses, which tested the relationship between ASVAB scores and course grades more directly. While ASVAB's Electronics Composite score demonstrated a moderate ability to predict success in training for white male students, it was less successful for female or minority students. The factor score we derived from ASVAB was in most cases the best simple predictor of training success because it utilized information from all ten ASVAB subtests, and not simply from the subset used for AFQT or the Electronics Composite. However, all three ASVAB measures (AFQT, Electronics, and factor scores) in most cases proved to be relatively weak predictors of performance in training for minority students.

Correlations do not imply causality, nor does the lack of a correlation for a subsample indicate the location of a problem. From our analyses it is impossible to conclude either that ASVAB is a weaker measure of ability for some groups, or that some factor in classroom training contributes differentially to the success of different groups. Yet, as the youth pool shrinks and its demographic characteristics shift, the military will find itself turning more toward minority and female recruits. These groups, as we have seen, consistently score lower in the measures used to assign recruits to technical training and in our largest service are less likely to perform well. It will become increasingly incumbent on all services to optimize selection criteria for technical advanced individual training for women and minority groups, to provide compensatory training where needed, and to assure that no extraneous factors within the training environment interfere with the full development of a recruit's potential.

Field Measures of Training Effectiveness

Whatever criteria may exist to predict or to assess a recruit's performance in training, the ultimate criterion of training effectiveness is the recruit's performance on the job. Our third evaluation question addresses this issue: How well do the services' selection criteria and training evaluation measures predict success in high technology roles?

To answer this question, we attempted to locate individual field-performance data routinely collected by the services that could be linked to our ASVAB and classroom training data to serve as reliable and valid indicators of training effectiveness. And, although we were made aware of numerous post-training evaluation activities performed by the individual services, only the Army could provide us with individual performance measures. In this chapter, we will examine the quantitative relationship between these Army data and the other information we compiled. We will also discuss other evaluation mechanisms used by the services and suggest a potential alternative source of post-training evaluation measures.

Army

Skill Qualification Test

By Army regulation, a soldier's occupational specialty performance is tested within six months of completion of training and every year thereafter. These written tests are prepared by the sponsoring training site. They are administered under the direction of the Skill Qualification Test (SQT) directorate at Fort Eustis, Virginia, where the resulting data are stored.

Fort Eustis provided us with the SQT scores of all soldiers who took the SQT from 1985 to 1988 in the occupational specialties we had chosen for our sample. Summary statistics for these data are provided in appendix IV. We matched these scores, where possible, with ASVAB scores and classroom grades for each soldier included in our training site review.[1] Table 4.1 presents the scores of these soldiers summarized by demographic groups, together with the correlation coefficient estimating the relationship between SQT and the measures we examined in the previous chapter.

[1]For soldiers with multiple SQT scores during this period, we used only the first score.

Table 4.1: Correlation of SQT and Predictor Variables

Category	Mean	Number	Correlation with SQT			
			AFQT[a]	Electronics Composite[b]	Factor[c]	Grade[d]
Male	82.12	209				
Raw			0.21[f]	0.28[f]	0.36[f]	0.47[f]
Adjusted[e]			0.30[f]	0.41[f]		
Female	77.52	21				
Raw			−0.07	0.12	−0.03	−0.52[f]
Adjusted[e]			−0.10	0.19		
White	81.86	144				
Raw			0.21[f]	0.25[f]	0.32[f]	0.44[f]
Adjusted[e]			0.33[f]	0.40[f]		
Nonwhite	81.45	86				
Raw			−0.19	0.07	0.12	0.44[f]
Adjusted[e]			−0.22	0.10		
Total	81.70	230				
Raw			0.18[f]	0.28[f]	0.34[f]	0.43[f]
Adjusted[e]			0.26[f]	0.41[f]		

[a]AFQT = sum of subtest standard scores

[b]Electronics Composite = sum of subtest standard scores for Electronics Composite

[c]Factor = score from first factor from principal component analysis

[d]Grade = final course grade

[e]Adjusted = adjusted for restriction of range

[f]$p < .05$

For the total universe of soldiers the best simple predictor of SQT scores is final classroom grades, which explains 18.5 percent of the variation in SQT's. The AFQT and Electronics scores from ASVAB scores were also significantly related to SQT's for white males in our sample, but factor scores consistently outpredicted these composites. For females and for nonwhite soldiers, however, ASVAB scores were not positively related to future performance as measured by SQT. Most surprisingly, the grades scored by female students at the training site were inversely correlated with their SQT scores—that is, women with higher grades tended to score lower on SQT's, and vice versa.

The limited size of our sample, especially for female soldiers, makes it inappropriate to generalize without severe caveats. However, our analysis suggests that the traditional ASVAB scores may not be the best predictor of performance for the nontraditional—that is, the female or minority—soldier. This finding reinforces the concern we expressed in

the last chapter, that better predictors of success for these groups should be found. Any interpretation of the inverse relationship between grades and SQT's for women would be purely speculative, but this anomaly warrants further investigation.

Other Evaluation-Related Activities

Each Army training site includes an evaluation unit that performs regular process evaluations. These include classroom observations of instructors, annual meetings to review curricula, cyclical outreach programs to contact graduates of the school in the field and their supervisors, and occasional more intensive curriculum reviews called training effectiveness analyses.

Classroom observations are conducted on a regular basis by both master trainers and the training site internal evaluation unit. They are performed more frequently when instructors are new or have received less-than-satisfactory evaluations. Most of the observation reports that we reviewed, particularly those performed by the internal evaluation unit, were mainly concerned with administrative details. The most frequent criticism we encountered was that copies of the lesson plan and curriculum materials were not properly arranged and situated at an empty desk in the rear of the classroom for the observer.

Schoolhouse external evaluation units also conduct outreach programs during which members of the units travel to Army bases—where a large concentration of the training-site graduates are stationed—to collect information on the opinions of base staff about training quality. These reviews occur approximately every two or three years for the courses we reviewed, but they are not routinely scheduled. They are more frequently occasioned by indications from the field of training problems, and their frequency is also affected by travel-budget considerations.

More objective and formal training effectiveness analyses are performed when a new training course is introduced or when weapons system modifications prompt major changes in the curriculum. These analyses include written tests, hands-on tests, and interviews with soldiers and their supervisors. The most recent training effectiveness analysis for the courses we reviewed was conducted during the summer of 1987 and was prompted by changes to the Hawk missile system.

Navy

Sources of Individual Field Performance Data

We considered two possible sources of field performance information routinely collected by the Navy as measures of the effectiveness of the training courses in our sample: Level II surveys and Advancement in Rating Examinations. The Level II survey program was designed to collect information on the job performance of recent training-school graduates.[2] For each course, questionnaires were sent to the supervisors of graduates approximately six months after graduation, asking them to rate individual tasks performed within the specialty (as to their importance) and the adequacy of the level of training demonstrated by the course graduates. We found, however, that Level II surveys have been effectively abandoned by the Navy, and that none has been performed since at least 1986.

Advancement in Rating Examinations are multiple-choice tests administered to candidates for promotion who have already been certified as qualified by their commanding officers. Different tests are prepared for each promotion cycle, and their results are used to rank candidates. Because they are not standardized, and are not administered to all graduates, these tests, in the judgment of test developers and administrators, are "not a good source of training evaluation feedback." We concurred with this judgment.

Internal Review of Evaluation Practices

In 1986, the Chief of Naval Operations requested that the Naval Training Systems Center (NTSC) determine the current status of Navy training evaluation and provide recommendations for the future conduct of such operations. NTSC submitted three reports to the Chief of Naval Technical Training in 1988. They identified three central evaluation functions: Level II surveys, the Fleet Training Assessment Program (FLETAP), and the Training Assessment Survey Team (TAST). The TAST concept had only recently been established at the time of the NTSC report, and only two surveys had been completed under the program. These surveys were limited to new weapons systems and involved fleet visits to identify training deficiencies and requirements and any corrective actions that needed to be taken.

[2]The term derives from a classification of evaluation intensiveness established in 1981 by the Naval Education Training Command. Level I refers to unsolicited feedback to training sites concerning training adequacy, Level II to a questionnaire sent to the fleet, and Level III to an in-depth analysis of problems identified in lower level reviews.

FLETAP is currently a reactive system that attempts to identify training deficiencies through either direct input from the fleet or review of reports and other fleet materials. FLETAP is also responsible for performing Training Quality Reviews, which involve administering job performance tests to fleet personnel to measure adequacy of training. No such reviews have been completed. The FLETAP component responsible for the Pacific Fleet consists of five full-time staff positions, four of which were filled at the time of our visit there. Its Atlantic Fleet counterpart has four authorized staff positions, three of which were filled.

The NTSC report also identified numerous other nonformal or noncentralized evaluation and evaluation-related activities within the Navy's training community. However, NTSC found that the quality of current Navy classroom training cannot be readily ascertained for the vast majority of courses; that there is a general lack of technical evaluation/assessment skills; that current evaluation activities are fractionated, not comprehensive, and operating in an environment of obsolete instructions and unclear objectives. NTSC concluded that the fleet's mandate to provide useful data to the training community about the performance of its graduates needed to be enforced and that fleet evaluation activities should be upgraded and appropriately staffed. It also recommended that internal training appraisal responsibility be decentralized to the training site level and that independent external programs be reviewed for technical adequacy and integrated into an overall systematic approach.

In response to these reports, a three-person team has recently been established at the headquarters of the Chief of Naval Education and Training to review the NTSC proposals and recommend an integrated training appraisal program. No firm timetable has yet been established for the team's report, but they anticipate providing a proposal in the summer of 1990. We welcome this Navy effort, but we question whether this response will prove adequate in view of the severity and extensiveness of the problems NTSC has documented.

Air Force

Sources of Individual Field Performance Data

We considered sources of individual-level data for field performance of Air Force personnel equivalent to those we considered for the Navy—that is, promotion examinations and supervisory surveys. After interviewing Air Force personnel, however, we concluded that neither was appropriate for our purposes.

Unlike the Navy's Level II surveys, the Air Force supervisory surveys are still in use. They are conducted by the training sites' evaluation units for each training course at 2- to 3-year intervals. Questionnaires are sent to the supervisors of recent training graduates to determine how frequently they perform each of the major tasks for which they were trained, and how well they perform them. A summary training evaluation report is produced from these data identifying task-specific training deficiencies and/or unnecessary training. We were informed that the individual-level data collected by these surveys are not maintained by the training sites after their reports have been prepared. Therefore, no individual data exist that would allow us to perform analyses equivalent to those we performed using the Army SQT data.

Other Evaluation-Related Activities

Other training assessment procedures exist, including training quality reports, utilization and training workshops, and occupational survey reports. Training quality reports provide a means for supervisors of recent training-site graduates to report apparent deficiencies in a recruit's training. Like the Navy's FLETAP activities, these reports are part of a reactive evaluation process. A succession of training quality reports for a given course can lead to a complete course review. The other activities are more concerned with front-end analysis. Occupational survey reports on occupational specialties are prepared approximately every three to four years. They are based on questionnaires designed to define the major tasks performed by specialists and their relative frequency. Utilization and training workshops are held when the job requirements of an old occupational specialty change dramatically or when a new specialty is defined. Major command functional officers, training staff officers, and managers at the Air Force technical schools participate by examining data from occupational survey reports and identifying the specific training requirements of the specialty.

Alternative Data Sources: The Job Performance Measurement Project

A key impediment to establishing a field evaluation component of training assessment is the expense of developing, testing, and administering measures that validly and reliably measure actual performance. Since the early 1980's, a major effort to address these measurement issues has been under way under the direction of the Office of Accession Policy of the Office of the Assistant Secretary of Defense for Force Management and Personnel. Known as the Joint-Service Job Performance Measurement (JPM) project, the effort was initiated at the request of the Congress to validate ASVAB measures against actual performance in the field—instead of against training grades, which had been the sole criterion. The project was triggered by the discovery of the ASVAB misnorming in the late 1970's, which unintentionally allowed some 300,000 less qualified recruits into the services and resulted in field commanders' complaints of quality deterioration among their personnel. JPM, in other words, was directed toward testing the connection between the first and third points in our model: test data collected for selection and classification purposes at recruitment, and field performance data. JPM did not set out to establish a link between classroom performance and field performance.

JPM concluded that suitable measures of field performance did not exist, and undertook to develop them. Over several years, some highly reliable hands-on performance tests were developed and administered for 25 occupational specialties across the four services. Surrogates for hands-on testing were also developed, including more traditional job-knowledge tests and performance ratings. JPM concluded that AFQT reliably predicted differences in levels of actual field performance, and that these differences tended to persist through a recruit's enlistment. JPM, however, has not reported any analyses of sex- or race-related differences. Because of its ASVAB orientation, the project also has not addressed the issue of the classroom/field-performance connection.

JPM performance measures were expensive to develop and frequently costly to administer, and they therefore may not be suitable for more routine use as measures of training effectiveness. However, the investment made to develop these measures and their surrogates could prove more profitable if some of the measures developed and the lessons learned in the JPM effort were more widely applied to the development of realistic assessment procedures for training.

Summary and Conclusions

Our third evaluation question asked to what extent the services' selection criteria and training evaluation measures predict success in high technology roles. While we identified a multitude of evaluation-related activities in the three services, we nevertheless concluded that insufficient data existed for us to respond to this question. Army SQT data can be adapted for this purpose, but neither the Navy nor the Air Force routinely collects and maintains field performance data to evaluate individual-level training effectiveness.

Our analysis of Army SQT data was hindered by the limited size of the sample. We were able to derive some preliminary conclusions, however—namely, that classroom performance, as measured by SQT, is a moderately strong indicator of future field performance for males, but not for females, and that ASVAB can predict SQT's moderately well for white male recruits, but is apparently unrelated to SQT scores achieved by women and minorities. These ASVAB/SQT findings are consistent with the pattern of ASVAB/course-grade relationships we discussed in the previous chapter.

The lack of other objective, systematically collected field evaluation data renders meaningful evaluation of training effectiveness impossible. Decisionmakers—whether they are in the Congress, DOD, or the individual services—can only react to problems in the field after they have become apparent and have been identified as training-related. However, given the cost and complexity of today's military equipment, it is imperative that the services possess adequate evaluative data to monitor how well personnel are being prepared to use and maintain these weapons.

Summary, Recommendations, and Agency Comments and Our Response

Summary

Our report has addressed three evaluation questions:

- How has the aptitude of recruits for technologically sophisticated specialties changed since 1980?
- How useful are the data collected by the services before and during classroom training for selecting individuals for high technology roles and for evaluating the effectiveness of this training?
- How well do the services' selection criteria and training evaluation measures predict success in high technology roles?

To respond to these questions, we examined the three essential types of information that could be used to assess the effectiveness of military training: (1) data collected at entry to the military for selection and assignment to an occupational specialty, (2) data on classroom measures of performance during formal training, and (3) data on individual field performance. Our analysis has been set in the context of a recruit pool shifting toward a much higher representation of women and minorities.

To answer the first question, we examined ASVAB scores during the 1980's and found that (1) most gains in recruit quality occurred in the first half of the decade, (2) technical abilities of recruits have begun to decline, and (3) women and minorities continue to score lower on technical measures than white males. These findings suggest that an increased burden will be placed on the services' training establishments to assure the technical competence of their future graduates. The services' response may also need to include more demographically sensitive training and/or additional compensatory training to raise basic skill levels.

Our response to the second question involved an analysis of classroom grades from thirteen technical courses. Our findings indicated that (1) some deficiencies exist in the Army's computerized grading system; (2) during training women and minorities overcome their initially lower technical scores in the Navy and Air Force, but not in the Army; (3) classroom success appears more related to a general ability level as measured by ASVAB than to the Electronics Composite score currently in use, particularly for women; and (4) ASVAB's ability to predict classroom success for minorities is weak.

The last three findings are interrelated. Unlike the Army, in the Navy and Air Force, women entered training with significantly higher AFQT scores than men. In addition, the gap in AFQT scores between whites and nonwhites was twice as large for Army trainees as for their Navy and

Air Force counterparts. Based on these findings, we concluded that the services should consider developing a more general ASVAB derivative, such as our factor score, to assign women and minorities to technical training.

We found that there was insufficient evidence to attribute the weak relationship between ASVAB and course grades for women and minorities either to problems with ASVAB or to factors in the training environment. Yet, whatever its source, the relative inconsistency of the two measures exists and should be addressed by both the recruiting and training communities.

In response to the third question, we examined post-classroom measures of training effectiveness. We concluded that (1) only the Army routinely collects data on individual field performance useful for training evaluation purposes; (2) on the basis of these Army data, ASVAB scores are even weaker predictors of field performance for women and minorities than of classroom success; and (3) the Navy's training evaluation component is in need of more intense review and reform than it is currently receiving.

In summary, we found serious weaknesses or gaps at each of the data points required by the evaluation model posited in chapter 1. Of these, the most serious deficiency is the inability of the Air Force and Navy to base their evaluation of their selection procedures and classroom training in systematically collected, objective field performance data. Without the ability to test the "fit" of these data points with one another, the services are not able to maximize their training effectiveness, or even to estimate realistically how successful their training investment is in producing skilled operators and maintainers of today's—and tomorrow's—sophisticated weaponry.

Recommendations

We believe that evaluating the effectiveness of the training provided by the services is crucial if they are to meet the future challenges of changing recruit demographics and increasingly sophisticated weaponry. Therefore, we make the following recommendations for action at each of the three information collection points that we consider essential to adequate training evaluation: (1) that the Office of Force Management and Personnel direct the personnel research it coordinates among the individual services to identify more sensitive predictors of classroom performance for women and minority students from the ASVAB data it already possesses; (2) that the Secretary of the Army direct the Training

and Doctrine Command to review the classroom grading procedures identified within the report as deficient, for their accuracy, appropriateness, and reliability; (3) that the Secretary of the Navy establish a firm deadline for developing a training evaluation program and that he direct that the adequacy of current resources allocated to this effort be reexamined. Finally, we recommend that the Assistant Secretary of Defense for Force Management and Personnel review alternative measures of field performance already developed by the services under the Job Performance Measurement project for their potential applicability to training and on-the-job performance evaluation.

Our purpose in this study has been to review the ability of the services to monitor, evaluate, and (where necessary) adjust training to changes in the demographics and technical ability of the recruit pool and to the technical sophistication of weapons systems. Whatever changes in our military posture are occasioned by shifts in the nature of threats to our national security, we believe that accurate information relating to the recruit pool, to the effectiveness of military training, and to on-the-job performance will continue to be essential to the mission of our armed forces.

Agency Comments and Our Response

In its written response to a draft of this report, DOD concurred with all of its recommendations and identified specific actions to be taken toward implementing them. DOD also concurred or partially concurred with what it identified as the main findings contained in the report. DOD also raised some technical methodological questions and offered some thoughtful interpretations of our findings. (See appendix V.) We have reviewed these comments and, where appropriate, have made changes to the text.

DOD generally agreed with our description of changes in recruits' ASVAB scores during the past decade. It commented, however, that it would be inappropriate to define a recruit's technological sophistication merely as his or her Electronics Composite score. We agree that this would be a very limited definition, and for this reason our report encouraged the development of better predictors of success in more technologically demanding occupational specialties. DOD's speculation that the decline in Electronics Information scores is attributable to a decline in technical vocational education in high schools is persuasive. It could as well have speculated that the lower Electronics Composite scores of women recruits are attributable to their traditionally lower enrollment in such courses.

DOD generally concurred with our analysis of classroom grades and their relationship to ASVAB predictors. However, it questioned the appropriateness of some of our procedures. DOD summarized its methodological concerns as (1) inappropriate pooling of grades from courses with different metrics, (2) implausibly high factor scores after correction for restriction in range, (3) lack of detailed regression analyses for differences between subgroups, and (4) small sample sizes for subgroups.

DOD incorrectly assumes that we simply pooled raw course grades from different courses. Before performing correlation analyses, we standardized course grades to a common metric to adjust for any differences between courses in grading procedures. We have also added to the draft we provided DOD parallel tables of results on the individual-course level. (See appendixes II and III.)

We share DOD's concern about the apparently inflated values of the adjusted validity coefficients for factor scores, but we disagree with their speculation that inappropriate statistical procedures are the source of this inflation. We applied the same conventional adjustment procedures to all three scores—AFQT, Electronics Composite, and factor scores—and, as DOD comments, for the first two scores our results "are consistent with other analyses." As we stated in the draft report, the factor scores were based on the ASVAB norm group correlation matrix provided us by DOD. Having performed a principal-components analysis of these data, we applied the resultant scoring coefficients to our sample to obtain factor scores. This procedure ideally offers two advantages. First, it bases the correlation analysis on a norm group presumably closer to the universe of applicants to military service than our sample of relatively high-scoring recruits. Second, it permits adjustment for restriction of range.

After thorough reexamination of our procedures and the data to which they were applied, we concluded that the results of factor analysis of the DOD correlation matrix should not be applied to our sample because of differences between the two samples in the magnitude of subtest intercorrelations. DOD reported substantially higher intercorrelations than were present in our sample. As a result, the variance of our sample's factor scores, when based on the DOD correlations, was inappropriately restricted, and the adjustment for range restriction was overestimated. (All other things being equal, the smaller the sample variance, the greater the adjustment for restriction in range.)

We therefore have recalculated our factor scores, deriving them from a principal-component analysis of our sample's ASVAB scores rather than from an analysis of the norm-group correlation matrix provided by DOD. Consequently, no adjustment for restriction of range would be appropriate for these scores. While the correlations of these factor scores with our criterion measures vary somewhat from those originally reported (being in some cases higher and in others lower), the slight differences in no way affect the conclusion that we reached in the draft report and with which DOD has agreed in both written and oral comments—namely, that a broader-based measure than the simple composites currently in use would provide a valuable predictor of classroom performance.

DOD cites the absence of certain regression-related statistics—intercepts, regression coefficients, and standard errors of estimates—and the small sample size in some subgroups as reasons for not "generalizing to other samples" or "making policy decisions" on the basis of our report. First, for simple bivariate relationships such as we analyzed (ASVAB versus course grades or SQT), our detailed reporting of means, N's, correlation coefficients, and significance levels serves essentially the same function as these equivalent regression statistics. We would, however, gladly provide our data base to DOD for alternative analysis. Second, we repeatedly draw the reader's attention to the problem of small sample size in some subgroups. Most importantly, we strongly agree that, unless they are replicated on larger samples, our analyses should not be the basis for significant policy shifts in selection and classification of recruits. Rather, we recommended (and DOD concurred) that the services attempt to develop more sensitive predictors of training success for minorities and women. (Indeed, one of the main strengths of our work here is that it determined the insensitivity to these populations of current predictors.) Should the results of these efforts prove successful, policy changes would then be appropriate.

The Army found "neither surprising nor particularly disturbing" the fact that we were not able to use many of the test scores they provided for some courses because they do not discriminate among soldiers' performances. We would point out that (1) the same software and report formats are used to assign scores to trainees in these courses as in other similar courses where we found usable scores; (2) we were able for some of these cases to reanalyze the individual measures and derive meaningful scores; and (3) the Army assigns and maintains rank-in-class statistics for each graduate of these courses on the basis of this software, thus itself implicitly measuring and recording the relative performance of individuals. While our ability to perform correlational analyses may

not be a critical need, in our opinion the Army's ability to perform objective evaluations of the effectiveness of its courses is. We therefore welcome the concurrence of the Army in our recommendation to review its testing procedures for the courses we identified.

DOD commented on our review of field measures of training effectiveness for each of the services, asserting that our negative view of ASVAB scores as a predictor of performance for female and minority soldiers was contrary to research on predicting training success. Not only does DOD provide no specifics on this research but also, and more importantly, it is not clear how predicting training outcomes is directly relevant to the issue of field performance. Of more interest are the preliminary results reported from ongoing research by the Army Research Institute. These results suggest a fairly strong relationship for women and a somewhat weaker, but still significant, relationship for blacks between ASVAB and SQT in larger occupational specialties. The Army appears to concede that these results may not be true for smaller, more technical specialties, such as the ones we examined. What is most noteworthy about the Army's response, however, is its capability to perform these analyses of field performance routinely, a capability that the Navy and Air Force do not share.

The Navy supplied some information on recent steps being taken to enhance training evaluation methods in addition to the ones we identified in the report. The Air Force commented that they do not have SQT's and do not plan to introduce them in the near future. It noted that "testing, recoding, and documenting individual performance for statistics is very time-consuming, requires additional manpower, and is cost-prohibitive." It would be difficult to agree with the Air Force that determining the effectiveness of individual performance is merely a statistical endeavor, or even that it is an optional one. Rather, it lies at the core of our ability to know how well we are prepared for meeting critical defense challenges. Indeed, given the cost and complexity of today's military equipment, it is imperative that all the services possess adequate evaluative data to monitor how well personnel are being trained to use and maintain these weapons. Our report does not propose the introduction of SQT's into other services, nor does it attempt to determine the cost-effectiveness of SQT's. It does, however, assert the need for objective, systematically collected information on individual field performance in all services.

Finally, DOD noted that it had directly addressed the applicability of lessons learned from the Joint-Service Job Performance Measurement Program in 1985, but had deferred implementing any training-related application of these measures at that time. DOD states that it will explore the feasibility of such an application once again.

AFQT Mean Score and Electronics Composite Summary Statistics: 1981-89

Table I.1: AFQT Mean Scores, by Gender[a]

	Male		Female	
Year	Number	Mean	Number	Mean
1981	163,571	203.95	22,886	202.95
1982	222,726	206.26	30,311	209.10
1983	227,161	209.51	32,546	211.57
1984	226,975	210.36	32,026	211.15
1985	222,772	211.55	35,368	211.43
1986	254,030	211.94	37,175	212.73
1987	239,122	212.17	35,385	212.42
1988	213,493	212.64	32,682	212.04
1989	217,783	211.83	35,984	211.78

[a]Sum of subtest standard scores

Table I.2: AFQT Mean Scores, by Service[a]

	Army		Navy		Air Force		Marine Corps	
Year	Number	Mean	Number	Mean	Number	Mean	Number	Mean
1981	76,284	195.52	47,715	208.61	37,389	213.12	25,069	206.16
1982	108,063	201.73	55,182	210.06	57,442	212.86	32,350	205.84
1983	121,112	206.07	55,256	212.52	51,771	216.72	31,568	207.78
1984	118,287	207.07	57,214	211.85	50,235	218.45	33,265	207.67
1985	111,625	209.30	59,604	211.92	57,617	217.08	29,294	208.34
1986	125,918	210.33	68,891	210.30	62,372	217.08	34,024	211.44
1987	120,538	210.73	66,078	210.75	54,371	218.10	33,520	210.90
1988	102,709	210.88	69,080	211.58	40,087	219.94	34,299	210.93
1989	106,126	209.42	73,272	210.40	42,247	220.59	32,122	211.45

[a]Sum of subtest standard scores

Table I.3: AFQT Mean Scores, by Race/Ethnicity[a]

	White		Black		Hispanic		Other	
Year	Number	Mean	Number	Mean	Number	Mean	Number	Number
1981	138,431	209.27	35,666	186.56	6,904	191.00	5,456	194.95
1982	189,134	211.48	48,377	190.86	8,569	193.97	6,957	198.91
1983	196,585	214.19	47,540	194.54	8,616	198.71	6,966	202.54
1984	193,193	215.07	48,500	194.99	9,439	199.46	7,869	204.15
1985	190,243	215.79	49,663	197.97	9,504	202.32	8,730	205.88
1986	212,661	215.94	56,150	199.20	12,059	204.26	10,335	206.74
1987	198,130	216.62	54,166	198.67	13,708	205.00	8,503	207.42
1988	174,501	217.16	50,370	199.14	13,567	205.92	7,737	207.84
1989	177,111	216.40	53,409	199.07	15,499	205.92	7,748	206.97

[a]Sum of subtest standard scores

Table I.4: AFQT Mean Score Overall Totals[a]

	Overall total	
Year	Number	Mean[b]
1981	186,457	203.83
1982	253,037	206.60
1983	259,707	209.77
1984	259,001	210.41
1985	258,140	211.53
1986	291,205	211.90
1987	274,507	212.21
1988	246,175	212.56
1989	253,767	211.82

[a]Sum of subtest standard scores

[b]Standard deviation = 20.66

Table I.5: Electronics Composite Mean Scores, by Gender[a]

	Male		Female	
Year	Number	Mean	Number	Mean
1981	163,571	207.89	22,886	194.41
1982	222,726	210.00	30,311	199.18
1983	227,161	212.91	32,546	201.52
1984	226,975	213.46	32,026	201.40
1985	222,772	212.70	35,368	199.57
1986	254,030	211.76	37,175	200.57
1987	239,122	212.17	35,385	200.57
1988	213,493	212.73	32,682	199.43
1989	217,783	211.50	35,984	199.97

[a]Sum of subtest standard scores

Table I.6: Electronics Composite Mean Scores, by Service[a]

	Army		Navy		Air Force		Marine Corps	
Year	Number	Mean	Number	Mean	Number	Mean	Number	Mean
1981	76,284	198.22	47,715	209.76	37,389	215.75	25,069	208.27
1982	108,063	204.03	55,182	210.33	57,442	215.24	32,350	207.90
1983	121,112	207.92	55,256	212.16	51,771	218.34	31,568	210.00
1984	118,287	208.56	57,214	211.69	50,235	219.87	33,265	209.70
1985	111,625	208.66	59,604	209.66	57,617	216.77	29,294	208.17
1986	125,918	208.73	68,891	207.32	62,372	215.48	34,024	209.80
1987	120,538	208.79	66,078	208.55	54,371	217.21	33,520	209.36
1988	102,709	209.11	69,080	208.71	40,087	219.01	34,299	209.53
1989	106,126	207.19	73,272	207.29	42,247	218.69	32,122	209.65

[a]Sum of subtest standard scores

Table I.7: Electronics Composite Mean Scores, by Race/Ethnicity[a]

	White		Black		Hispanic		Other	
Year	Number	Mean	Number	Mean	Number	Mean	Number	Mean
1981	138,431	212.47	35,666	186.45	6,904	193.40	5,456	197.91
1982	189,134	214.51	48,377	190.01	8,569	196.37	6,957	201.33
1983	196,585	216.81	47,540	193.24	8,616	200.93	6,966	204.31
1984	193,193	217.53	48,500	193.49	9,439	201.35	7,869	206.24
1985	190,243	216.28	49,663	193.94	9,504	202.50	8,730	205.87
1986	212,661	215.50	56,150	194.11	12,059	203.07	10,335	205.78
1987	198,130	216.19	54,166	193.50	13,708	203.76	8,503	207.23
1988	174,501	216.86	50,370	194.08	13,567	204.54	7,737	207.08
1989	177,111	215.64	53,409	193.46	15,199	203.66	7,748	206.57

[a]Sum of subtest standard scores

Table I.8: Electronics Composite Mean Score Overall Totals[a]

	Overall total	
Year	Number	Mean[b]
1981	186,457	206.04
1982	253,037	208.44
1983	259,707	211.15
1984	259,001	211.59
1985	258,140	210.65
1986	291,205	209.97
1987	274,507	210.47
1988	246,175	210.67
1989	253,767	209.45

[a]Sum of subtest standard scores

[b]Standard deviation = 22.19

Predictor and Criterion Variable Mean Scores

Table II.1: Army Mean Scores

Category	AFQT[a]		Electronics Composite[a]		Course grade		SQT[b]	
	Mean	Number	Mean	Number	Mean	Number	Mean	Number
24J	227.87	65	234.75	65	86.75	76	82.58	53
27N	226.73	100	232.85	100	88.78	138	83.95	110
29V	238.22	136	242.92	136	93.55	41	76.98	65
Male	232.14	280	238.46	280	89.23	232	82.12	209
Female	232.87	23	230.13	23	80.31	23	77.52	21
White	234.00	255	240.00	255	90.19	160	81.86	144
Nonwhite	222.67	48	226.29	48	86.86	95	81.45	86
All Army	232.20	303	237.83	303	88.94	255	81.70	230

[a]Sum of subtest standard scores

[b]Score on Skills Qualification Test

Table II.2: Navy Mean Scores

Category	AFQT[a]		Electronics Composite[a]		Course grade	
	Mean	Number	Mean	Number	Mean	Number
AQ	228.10	783	233.13	783	89.72	833
AX	231.64	392	236.16	392	90.64	469
STG	228.57	3,233	234.43	3,233	90.23	3,418
STS	231.87	1,698	237.47	1,698	86.89	1,723
Male	229.59	6,080	235.33	6,080	89.11	5,882
Female	235.59	76	230.65	76	90.70	71
White	230.49	5,355	236.25	5,355	89.20	5,179
Nonwhite	224.18	801	228.74	801	89.57	1,159
All Navy	229.67	6,156	235.27	6,156	89.30	6,443

[a]Sum of subtest standard scores

Table II.3: Air Force Mean Scores

Category	AFQT[a]		Electronics Composite[a]		Course grade	
	Mean	Number	Mean	Number	Mean	Number
45530A	235.53	119	240.72	119	90.17	119
45530B	235.93	231	240.55	231	90.82	231
30332	238.12	212	245.00	212	91.77	227
30333	234.15	360	239.77	360	91.31	377
Male	235.45	824	241.94	824	91.31	854
Female	237.73	98	235.88	98	89.91	100
White	236.22	825	241.95	825	91.21	855
Nonwhite	231.19	97	235.73	97	90.76	90
All Air Force	235.68	922	241.29	922	91.16	954

[a]Sum of subtest standard scores

Intercorrelation of Study Variables by Occupational Specialty

Table III.1: Intercorrelation of Study Variables: Army, 24J[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1 00	0 79[g]	0 83[g]	0 31[g]	0 49[g]
Electronics Composite	65	1 00	0 81[g]	0 32[g]	0 33[g]
Factor	65	65	1 00	0 40[g]	
Grade	59	59	59	1 00	
Male					
AFQT	1 00	0 82[g]	0 85[g]	0 29[g]	0 47[g]
Electronics Composite	55	1 00	0 79[g]	0 28[g]	0 30[g]
Factor	55	55	1 00	0 38[g]	
Grade	50	50	50	1 00	
Female					
AFQT	1 00	0 81[g]	0 89[g]	0 43	0 63
Electronics Composite	10	1 00	0 88[g]	0 15	0 15
Factor	10	10	1 00	0 21	
Grade	9	9	9	1 00	
White					
AFQT	1 00	0 82[g]	0 80[g]	0 24	0 39
Electronics Composite	49	1 00	0 79[g]	0 27	0 29
Factor	49	49	1 00	0 42[g]	
Grade	44	44	44	1 00	
Nonwhite					
AFQT	1 00	0 61[g]	0 80[g]	0 13	0 23
Electronics Composite	16	1 00	0 84[g]	0 15	0 16
Factor	16	16	1 00	0 17	
Grade	15	15	15	1 00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Table III.2: Intercorrelation of Study Variables: Army, 27N[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.84[g]	0.85[g]	0.36[g]	0.55[g]
Electronics Composite	100	1.00	0.92[g]	0.53[g]	0.57[g]
Factor	100	100	1.00	0.48[g]	
Grade	95	95	95	1.00	
Male					
AFQT	1.00	0.86[g]	0.85[g]	0.39[g]	0.59[g]
Electronics Composite	94	1.00	0.93[g]	0.52[g]	0.56[g]
Factor	94	94	1.00	0.48[g]	
Grade	89	89	89	1.00	
Female					
AFQT	1.00	0.86[g]	0.82[g]	0.84[g]	0.94[g]
Electronics Composite	6	1.00	0.96[g]	0.88[g]	0.93[g]
Factor	6	6	1.00	0.90[g]	
Grade	6	6	6	1.00	
White					
AFQT	1.00	0.82[g]	0.82[g]	0.31[g]	0.49[g]
Electronics Composite	85	1.00	0.90[g]	0.49[g]	0.52[g]
Factor	85	85	1.00	0.43[g]	
Grade	81	81	81	1.00	
Nonwhite					
AFQT	1.00	0.80[g]	0.81[g]	0.31	0.49
Electronics Composite	15	1.00	0.93[g]	0.65[g]	0.69[g]
Factor	15	15	1.00	0.62[g]	
Grade	14	14	14	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Table III.3: Intercorrelation of Study Variables: Army, 29V[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1 00	0 74[g]	0 79[g]	0 20	0 33
Electronics Composite	136	1 00	0 88[g]	0 50[g]	0 53[g]
Factor	136	136	1 00	0 38[g]	
Grade	35	35	35	1 00	
Male					
AFQT	1 00	0 75[g]	0 80[g]	0 25	0 41
Electronics Composite	129	1 00	0 88[g]	0 47[g]	0 50[g]
Factor	129	129	1 00	0 36[g]	
Grade	32	32	32	1 00	
Female					
AFQT	1 00	0 83[g]	0 80[g]	0 59	0 78
Electronics Composite	7	1 00	0 90[g]	0 79	0 84
Factor	7	7	1 00	0 57	
Grade	3	3	3	1 00	
White					
AFQT	1 00	0 74[g]	0 78[g]	0 20	0 33
Electronics Composite	119	1 00	0 87[g]	0 53[g]	0 56[g]
Factor	119	119	1 00	0 40[g]	
Grade	29	29	29	1 00	
Nonwhite					
AFQT	1 00	0 76[g]	0 85[g]	0 18	0 31
Electronics Composite	17	1 00	0 86[g]	0 34	0 36
Factor	17	17	1 00	0 23	
Grade	6	6	6	1 00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Table III.4: Intercorrelation of Study Variables: Navy, AQ[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.83[g]	0.85[g]	0.25[g]	0.40[g]
Electronics Composite	783	1.00	0.86[g]	0.27[g]	0.29[g]
Factor	783	783	1.00	0.25[g]	
Grade	774	774	774	1.00	
Male[h]					
AFQT	1.00	0.83[g]	0.85[g]	0.25[g]	0.40[g]
Electronics Composite	783	1.00	0.86[g]	0.27[g]	0.29[g]
Factor	783	783	1.00	0.25[g]	
Grade	774	774	774	1.00	
White					
AFQT	1.00	0.83[g]	0.84[g]	0.25[g]	0.41[g]
Electronics Composite	665	1.00	0.86[g]	0.28[g]	0.30[g]
Factor	665	665	1.00	0.27[g]	
Grade	656	656	656	1.00	
Nonwhite					
AFQT	1.00	0.82[g]	0.86[g]	0.13	0.22
Electronics Composite	118	1.00	0.83[g]	0.16	0.17
Factor	118	118	1.00	0.07	
Grade	118	118	118	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal.

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

[h]Women are prohibited from serving in the Navy's AQ occupational specialty

Table III.5: Intercorrelation of Study Variables: Navy, AX[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.81[g]	0.83[g]	0.41[g]	0.61[g]
Electronics Composite	392	1.00	0.89[g]	0.40[g]	0.43[g]
Factor	392	392	1.00	0.39[g]	
Grade	391	391	391	1.00	
Male					
AFQT	1.00	0.87[g]	0.88[g]	0.42[g]	0.62[g]
Electronics Composite	321	1.00	0.90[g]	0.43[g]	0.46[g]
Factor	321	321	1.00	0.41[g]	
Grade	320	320	320	1.00	
Female					
AFQT	1.00	0.75[g]	0.80[g]	0.39[g]	0.58[g]
Electronics Composite	71	1.00	0.83[g]	0.32[g]	0.34[g]
Factor	71	71	1.00	0.39[g]	
Grade	71	71	71	1.00	
White					
AFQT	1.00	0.80[g]	0.83[g]	0.44[g]	0.65[g]
Electronics Composite	336	1.00	0.89[g]	0.46[g]	0.49[g]
Factor	336	336	1.00	0.44[g]	
Grade	335	335	335	1.00	
Nonwhite					
AFQT	1.00	0.78[g]	0.84[g]	0.18	0.29
Electronics Composite	56	1.00	0.87[g]	0.02	0.02
Factor	56	56	1.00	0.07	
Grade	56	56	56	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal.

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Table III.6: Intercorrelation of Study Variables: Navy, STG[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.78[g]	0.80[g]	0.30[g]	0.48[g]
Electronics Composite	3233	1.00	0.84[g]	0.26[g]	0.28[g]
Factor	3233	3233	1.00	0.28[g]	
Grade	3123	3123	3123	1.00	
Male[h]					
AFQT	1.00	0.78[g]	0.80[g]	0.30[g]	0.48[g]
Electronics Composite	3233	1.00	0.84[g]	0.26[g]	0.28[g]
Factor	3233	3233	1.00	0.28[g]	
Grade	3123	3123	3123	1.00	
White					
AFQT	1.00	0.79[g]	0.80[g]	0.31[g]	0.49[g]
Electronics Composite	2791	1.00	0.84[g]	0.28[g]	0.29[g]
Factor	2791	2791	1.00	0.30[g]	
Grade	2697	2697	2697	1.00	
Nonwhite					
AFQT	1.00	0.71[g]	0.76[g]	0.22[g]	0.37[g]
Electronics Composite	442	1.00	0.78[g]	0.16[g]	0.16[g]
Factor	442	442	1.00	0.12[g]	
Grade	426	426	426	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

[h]Women are prohibited from serving in the Navy's STG occupational specialty

Table III.7: Intercorrelation of Study Variables: Navy, STS[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.76[g]	0.78[g]	0.28[g]	0.45[g]
Electronics Composite	1698	1.00	0.85[g]	0.26[g]	0.27[g]
Factor	1698	1698	1.00	0.26[g]	
Grade	1651	1651	1651	1.00	
Male[h]					
AFQT	1.00	0.76[g]	0.78[g]	0.28[g]	0.45[g]
Electronics Composite	1698	1.00	0.85[g]	0.26[g]	0.27[g]
Factor	1698	1698	1.00	0.26[g]	
Grade	1651	1651	1651	1.00	
White					
AFQT	1.00	0.77[g]	0.79[g]	0.28[g]	0.46[g]
Electronics Composite	1518	1.00	0.85[g]	0.27[g]	0.29[g]
Factor	1518	1518	1.00	0.28[g]	
Grade	1477	1477	1477	1.00	
Nonwhite					
AFQT	1.00	0.70[g]	0.68[g]	0.27[g]	0.44[g]
Electronics Composite	180	1.00	0.82[g]	0.11	0.12
Factor	180	180	1.00	0.12	
Grade	174	174	174	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal.

[b]AFQT = sum of subtest standard scores.

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite.

[d]Factor = score from first factor from principal component analysis.

[e]Grade = final course grade.

[f]Adjusted = correlation adjusted for restriction of range.

[g]$p < .05$

[h]Women are prohibited from serving in the Navy's STS occupational specialty.

Table III.8: Intercorrelation of Study Variables: Air Force, 45530A[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.74[g]	0.19[g]	0.22[g]	0.36[g]
Electronics Composite	119	1.00	0.87	0.27[g]	0.29[g]
Factor	119	119	1.00	0.30[g]	
Grade	119	119	119	1.00	
Male					
AFQT	1.00	0.77[g]	0.77[g]	0.21[g]	0.35[g]
Electronics Composite	99	1.00	0.86[g]	0.26[g]	0.28[g]
Factor	99	99	1.00	0.27[g]	
Grade	99	99	99	1.00	
Female					
AFQT	1.00	0.69[g]	0.63[g]	0.31	0.49
Electronics Composite	20	1.00	0.84[g]	0.15	0.15
Factor	20	20	1.00	0.25	
Grade	20	20	20	1.00	
White					
AFQT	1.00	0.75[g]	0.73[g]	0.24[g]	0.39[g]
Electronics Composite	102	1.00	0.87[g]	0.28[g]	0.29[g]
Factor	102	102	1.00	0.28[g]	
Grade	102	102	2102	1.00	
Nonwhite					
AFQT	1.00	0.58[g]	0.65[g]	0.08	0.13
Electronics Composite	17	1.00	0.85[g]	0.22	0.23
Factor	17	17	1.00	0.33	
Grade	17	17	17	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Table III.9: Intercorrelation of Study Variables: Air Force, 45530B[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.70[g]	0.72[g]	0.22[g]	0.36[g]
Electronics Composite	231	1.00	0.83[g]	0.27[g]	0.28[g]
Factor	231	231	1.00	0.29[g]	
Grade	231	231	231	1.00	
Male					
AFQT	1.00	0.71[g]	0.72[g]	0.23[g]	0.37[g]
Electronics Composite	215	1.00	0.84[g]	0.25[g]	0.27[g]
Factor	215	215	1.00	0.29[g]	
Grade	215	215	215	1.00	
Female					
AFQT	1.00	0.81[g]	0.83[g]	0.15	0.26
Electronics Composite	16	1.00	0.71[g]	0.25	0.26
Factor	16	16	1.00	0.10	
Grade	16	16	16	1.00	
White					
AFQT	1.00	0.70[g]	0.72[g]	0.25[g]	0.40[g]
Electronics Composite	206	1.00	0.81[g]	0.32[g]	0.34[g]
Factor	206	206	1.00	0.35[g]	
Grade	206	206	206	1.00	
Nonwhite					
AFQT	1.00	0.66[g]	0.65[g]	0.11	0.19
Electronics Composite	25	1.00	0.90[g]	0.05	0.06
Factor	25	25	1.00	0.04	
Grade	25	25	25	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal.

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Table III.10: Intercorrelation of Study Variables: Air Force, 30332[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.69[g]	0.75[g]	0.39[g]	0.59[g]
Electronics Composite	212	1.00	0.81[g]	0.41[g]	0.43[g]
Factor	212	212	1.00	0.43[g]	
Grade	212	212	212	1.00	
Male					
AFQT	1.00	0.74[g]	0.78[g]	0.41[g]	0.61[g]
Electronics Composite	186	1.00	0.82[g]	0.40[g]	0.42[g]
Factor	186	186	1.00	0.45[g]	
Grade	186	186	186	1.00	
Female					
AFQT	1.00	0.62[g]	0.71[g]	0.34	0.53
Electronics Composite	26	1.00	0.79[g]	0.48[g]	0.50[g]
Factor	26	26	1.00	0.31	
Grade	26	26	26	1.00	
White					
AFQT	1.00	0.70[g]	0.77[g]	0.36[g]	0.55[g]
Electronics Composite	190	1.00	0.81[g]	0.41[g]	0.43[g]
Factor	190	190	1.00	0.42[g]	
Grade	190	190	190	1.00	
Nonwhite					
AFQT	1.00	0.56[g]	0.70[g]	0.62[g]	0.81[g]
Electronics Composite	22	1.00	0.75[g]	0.43[g]	0.46[g]
Factor	22	22	1.00	0.61[g]	
Grade	22	22	22	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Table III.11: Intercorrelation of Study Variables: Air Force, 30333[a]

Category	AFQT[b]	Electronics Composite[c]	Factor[d]	Grade[e] Raw	Grade[e] Adjusted[f]
Total					
AFQT	1.00	0.72[g]	0.77[g]	0.32[g]	0.50[g]
Electronics Composite	360	1.00	0.83[g]	0.38[g]	0.40[g]
Factor	360	360	1.00	0.40[g]	
Grade	360	360	360	1.00	
Male					
AFQT	1.00	0.75[g]	0.79[g]	0.31[g]	0.49[g]
Electronics Composite	324	1.00	0.84[g]	0.39[g]	0.41[g]
Factor	324	324	1.00	0.34[g]	
Grade	324	324	324	1.00	
Female					
AFQT	1.00	0.58[g]	0.78[g]	0.50[g]	0.70[g]
Electronics Composite	36	1.00	0.74[g]	0.22	0.24
Factor	36	36	1.00	0.36[g]	
Grade	36	36	36	1.00	
White					
AFQT	1.00	0.71[g]	0.77[g]	0.34[g]	0.53[g]
Electronics Composite	327	1.00	0.84[g]	0.38[g]	0.40[g]
Factor	327	327	1.00	0.35[g]	
Grade	327	327	327	1.00	
Nonwhite					
AFQT	1.00	0.66[g]	0.68[g]	0.10	0.17
Electronics Composite	33	1.00	0.70[g]	0.43[g]	0.46[g]
Factor	33	33	1.00	0.43[g]	
Grade	33	33	33	1.00	

[a]Correlation coefficients are in upper diagonal and number in lower diagonal

[b]AFQT = sum of subtest standard scores

[c]Electronics Composite = sum of subtest standard scores for Electronics Composite

[d]Factor = score from first factor from principal component analysis

[e]Grade = final course grade

[f]Adjusted = correlation adjusted for restriction of range

[g]$p < .05$

Army SQT Mean Scores, by Occupational Specialty

Specialty	Year	Number	Mean
24J	1985	154	86.48
	1986	152	87.11
	1987	102	82.50
	1988	92	83.05
	Total	**500**	**85.23**
27N	1985	196	85.53
	1986	157	88.36
	1987	145	86.66
	1988	185	79.56
	Total	**683**	**84.81**
26V/29V	1985	1,308	82.28
	1986	1,261	79.39
	1987	944	80.19
	1988	831	78.77
	Total	**4,344**	**80.40**

Comments From the Department of Defense

ASSISTANT SECRETARY OF DEFENSE

WASHINGTON, D.C. 20301-4000

FORCE MANAGEMENT AND PERSONNEL

10 AUG 1990

Ms. Eleanor Chelimsky
Assistant Comptroller General
Program Evaluation and Methodology Division
U.S. General Accounting Office
441 G. Street, NW
Washington, DC 20548

Dear Ms. Chelimsky:

This is the Department of Defense (DoD) response to the General Accounting Office (GAO) draft report, "MILITARY TRAINING: Effectiveness for Technical Specialties Inadequately Measured," dated May 31, 1990 (GAO Code 973276, OSD Case 8371).

The report provides a series of useful recommendations that are consistent with ongoing DoD initiatives designed to develop more sensitive indicators of trainee performance and to develop more cost-effective ways of measuring performance both in the schoolhouse and on-the-job. Despite general agreement with the report's final recommendations, the DoD does not fully concur with many of the specific findings. In several cases, the findings and conclusions appear to be based on incorrect assumptions or inappropriate methodology. Specific issues and details are provided in the enclosure.

In addition, it is important to note that the field of job performance measurement is still a developing science and cost-effective measures for use in evaluating training effectiveness are not yet available. As discussed in the enclosure, the DoD has additional measurement programs in place beyond those discussed in the report, and continues to support a substantial number of research efforts to expand the boundaries of this science. The GAO report substantiates the Department's conclusions about the demands of selecting and training individuals to meet the requirements of technical specialties in the coming years, and reinforces current DoD efforts in this area.

The DoD appreciates the opportunity to comment on the draft report.

Sincerely,

Christopher Jehn

Enclosure:
As stated

80

**GAO DRAFT REPORT-DATED MAY 31, 1990
(GAO CODE 973276) OSD CASE 8371**

"MILITARY TRAINING: EFFECTIVENESS FOR TECHNICAL SPECIALTIES INADEQUATELY MEASURED"

DEPARTMENT OF DEFENSE COMMENTS

* * * * * *

FINDINGS

FINDING A: Background: Recruit Quality. The GAO reported that, if the entry level aptitude, knowledge, and skills of new recruits should fall short of human requirements needed to operate and maintain new technologically sophisticated weapons systems, greater demands would be placed on the Armed Services to compensate for the shortfall through training. The GAO observed that the recruit quality had grown in the eighties, as evidenced by the following statistics:

- in 1980, 68 percent of recruits were high school graduates, by 1986, 92 percent had high school diplomas; and
- in 1980, 65 percent of the recruits were in the top three mental categories on the Armed Forces Qualifying Test, compared with 96 percent in 1986.

The GAO also reported that:

- the number of young people available for the military recruit pool will continue to diminish until the mid-1990s;
- by the year 2000, five of every six new labor force entrants will be female, minority group members, or immigrants; and
- the graduates of the American educational system are said to be falling behind the youth of competitor nations in technological literacy--while, at the same time, weapons systems become increasingly sophisticated.

The GAO also reported that the Air Force has expressed concern about the quality of recruits, the Navy noted an erosion of its Delayed Entry Pool, and for the first time in 8 years, the Army failed to meet its quarterly recruiting quota in the first quarter of FY 1989. (pp. 1-1 to 1-5/GAO Draft Report)

DoD Response: Concur. While the statements attributed to the Services are essentially correct, they do not provide the "big picture." Since FY 1984, quality in the Air Force has remained stable at 98 to 99 percent high school diploma graduates and 98 to 100 percent individuals who score average or above on the enlistment test. Simultaneously, Air Force recruiting objectives have fallen from 60,000 in FY 1984 to 43,000 in FY 1989, making it easier to meet its goals with high quality. Although the Navy Delayed Entry Program pool eroded in FY 1989, it is back on target. And while the Army did not achieve it's first quarter FY 1989 recruiting objective (enlisting all but 475 of the 24,143 people it sought), it finished FY 1989 exceeding the objective. In addition, the impact of the mid-1990s dip in the size of the youth population will be moderated by reductions in accession requirements that are likely to be part of the overall down sizing of the military during this decade.

The GAO report also mentions that American youth are falling behind youth of competitor nations in "technological literacy." While unaware of the existence of international "technological literacy" data, it is the DoD objective to enlist those youth who can acquire the skills to field sophisticated weapon systems. To that end, the education of the nation's youth is of paramount importance to the DoD. Given students' lackluster performance on both national and international tests over the last decade, the DoD has formed a collaborative, working arrangement with the U.S. Department of Education, whereby the Department is assisting them with development and fielding of new international literacy tests. The DoD is also experimenting with those same tests with hopes of improving the Joint-Service enlistment test. The Department shares the GAO concern and hopes to have much-improved, international comparative literacy data over the next several years.

FINDING B: **The Quality of Military Recruits--1981-1989 Test Results**. The GAO reported that the Armed Services Vocational Aptitude Battery is comprised of ten subtests measuring abilities considered important for Military Service. The GAO also reported that all the Services use the same component subtests for two composite scores; the Electronics composite and the Armed Forces Qualification Test, which is the primary mental criteria for entry into the Armed Forces. The GAO found the following regarding Armed Forces Qualification Test:

- overall scores improved about 4 percent between 1981 and 1989;
- male recruit scores began and ended the decade slightly higher than female scores;

- scores differed more substantially across racial groupings than between genders;
- white recruits scores began the decade 10 percent higher than minority scores and ended 7 percent higher;
- mean scores for all Services were significantly higher in 1989 than 1981;
- Army scores began the decade substantially below those of the other Services, but by 1986, had reached the same level as Navy and Marine Corps recruits; and
- average Air Force scores have consistently been higher than the other Services and have not displayed their tendency to plateau at mid-decade levels.

The GAO found the following regarding the Electronics Composite:

- mean scores rose 2 percent between 1981 and 1989;
- scores peaked in 1984 and have shown a gradual decline since then;
- female recruits scored approximately 5 percent lower than male recruits during the eighties;
- white recruits scored about 11 percent higher than minorities in 1981 and 9 percent higher by 1989;
- the narrowing of the gap for minorities, however, was achieved in the first half of the decade--by 1989, scores for all racial groups were declining;
- the interservice pattern of scores mirror those of the Armed Forces Qualification Test, with the Army making up a 10 point difference with the Navy and Marines by 1986, and the Air Force on top throughout; and
- mean scores for the three Services changed very little from 1985 to 1988, but Army and Navy scores declined significantly in 1989. (pp. 2-1 to 2-7/GAO Draft Report)

DoD Response: Partially concur. Although the individual calculations have not been corroborated by the DoD due to time constraints, trends reported in the Armed Forces Qualification Test score data presented for comparison of groups (i.e., gender, race/ethnicity, and Service) look reasonable, as do the trends

reported regarding the Electronics Composite. Some technical questions suggest, however, that clarification may be necessary in the GAO narrative.

For example, the GAO report states that Armed Forces Qualification Test "scores improved about 4 percent between 1981 and 1989." In other statements, various percentage changes are mentioned for the Armed Forces Qualification Test and the Electronics Composite. Computing percentage gains or changes in subtest standard scores is not statistically appropriate. Scores on the Armed Services Vocational Aptitude Battery, of which the Armed Forces Qualification Test and the Composite scores are a part, do not have a meaningful zero point and, therefore, percentage changes cannot be interpreted. Computation of percentages requires a ratio scale, which is more powerful than the score scale for all aptitude tests, including the Armed Services Vocational Aptitude Battery. The same limitation applies to interpreting changes on the Electronics Composite.

Some factors related to changes in how scores have been computed are relevant, particularly since the report examines scores across several years. Between 1981 and 1989, there were several changes in the Armed Forces Qualification Test (e.g., the subtests used to compute the Armed Forces Qualification Test score were changed and the reference population for norming of the test was updated). It is unclear if the differences in how scores were computed over the years were taken into account in the analyses presented in Appendix 1 and Figures 1, 2, and 3; clarification as to these differences appears appropriate, otherwise comparisons of means will not be interpretable. The same sort of changes occurred over the years in the calculation of the Electronics Composite and would affect interpretation of Figures 5, 6, and 7.

Finally, with the large sample sizes achieved in the data analyses, statistical significance can be observed for differences that have relatively little practical significance. For example, while the statement that " . . . Navy scores declined significantly in 1989 (relative to 1988)" is true, the drop was from a score of 211.58 in 1988 to a score of 210.40 in 1989. That small a drop from one year to the next would be worth noting, yet not cause for alarm.

FINDING C: The Quality of Military Recruits--Number of Recruits Qualified for High Technology Specialties During the Period 1981-1989. The GAO reported that , as another measure of recruit qualification trends, it enumerated the number of recruits whose Armed Services Vocational Aptitude Battery scores met minimum standards required for entry into two selected high technology

military specialties: (1) air traffic controllers and (2) systems repair technicians. The GAO found the following for the air traffic controller specialty:

- in 1981, approximately 38,000 recruits qualified for the specialty and by 1986, more than 69,000 recruits qualified--but, since then, the number qualifying has declined to 58,000;
- in 1981, 87 percent of the qualifying recruits were white males, while two-thirds of all recruits were white males;
- by 1989, 84 percent of the qualifying recruits were white males, while only 61 percent of the recruits were white males
- while one third of the white males entering the Service qualified on the basis of their Electronics scores, fewer than 15 percent of the white females so qualified and fewer than 10 percent of the minority males and 3 percent of the minority females qualified on the basis of their Electronics scores.

The GAO found the following for the Systems Repair Technician:

- in 1981, the number of qualified recruits for the System Repair Technician specialty numbered 16,563 and, by 1983, the number had increased sharply--but by 1989, it had fallen back to within 700 of the 1981 level; and
- the vast majority of those qualified were white males, of whom 11 percent qualified compared with less than 2 percent for other demographic groups.

The GAO concluded that, based on its review, recruit quality trends during the eighties are not reassuring. The GAO also observed that fewer recruits are qualifying for the more demanding technical occupational specialties. The GAO further concluded that, with women and minorities forming the bulk of the new entry labor force by the year 2000, providing well-trained personnel for a technologically sophisticated military can be expected to become increasingly difficult. The GAO also noted that, in turn, the burden on training will increase, along with the need to monitor its effectiveness. (pp. 2-7 to 2-11/GAO Draft Report)

DoD Response: Partially concur. Providing well-trained personnel will become increasingly difficult should recruit quality

diminish. However, the DoD does not consider that recruit quality trends during the eighties, particularly the mid-to-late 1980s, are troublesome. During the last half of the decade, recruit quality has never been better. Compared to the youth population from which the DoD recruits, the quality level has consistently been well above average. For example, in FY 1989, 92 percent of new recruits had a high school diploma, in contrast to 74 percent in the youth population. Also, in FY 1989, 94 percent of new recruits scored average or above on the enlistment test, compared to 69 percent in the youth population.

Although it is reasonable that the GAO would want to assess how the aptitude of recruits for technologically sophisticated specialties has changed since 1980, the methodology selected to do so is flawed. Equating a decline on the Armed Services Vocational Aptitude Battery's electronics composite to a decline in recruits' "technological sophistication" is inappropriate. The electronics composite is composed of four subtests that measure mathematics ability (arithmetic reasoning and mathematics knowledge), general science, and electronics information. As the report Figure 8 indicates, the decline in performance on the composite is driven primarily by the decline in performance on one subtest--electronics information.

There is also a flaw in the example used by the GAO beginning on page 2-8, wherein the report refers to the Air Traffic Control specialty as having a minimum entry standard as of May 1989 of 230 on the Electronics composite (in standard score form). Air Traffic Control, Air Force Specialty Code 272X0, is selected on the General Composite and has never had an Electronics requirement. That renders report Figure 9 incorrect, if based on the composite described in the text. The GAO may have actually performed its analyses on the specialty titled Aircraft Control and Warning Radar Specialist, Air Force Specialty Code 303X2; in report Table 3.7, that specialty is correctly reflected as having an Electronics Composite qualifying score of 230.

The other specialty used by the GAO in this finding is Systems Repair Technician, an occupation so specialized that it is not assigned an Air Force Specialty Code, but is identified by a Reporting Identifier (99104). It would be appropriate for the report to mention that individuals qualifying for this specialty are not qualified for a "typical" high-technology job, but are at the very highest end of the technical continuum. A footnote identifying the specialty and its cutoff score requirement would be appropriate, similar to the footnote given at the bottom of page 2-8 for the other specialty.

It is speculated that the test score decline on the electronics information subtest is attributable to nationwide educational

curriculum changes. Over the course of this decade, dramatic changes have occurred in public and private elementary and secondary education programs.. These reforms have been well publicized and documented. As high school graduation standards have become more stringent, students have had fewer opportunities to take elective coursework. Consequently, enrollment in vocational education courses, like electronics/electricity, has declined dramatically. Throughout the 1980s, recruit quality, as measured on the Armed Services Vocational Aptitude Battery's Armed Forces Qualification Test composite, has improved. However, as the GAO pointed out, performance on the electronics subtest/composite has declined. Again, this is considered to be an artifact of the educational reform movement. Students simply are no longer enrolling in the technical and trade vocational classes where they can learn basic electronics/electrical constructs.

The electronics composite is a valid predictor of success in training and on the job for occupational specialties requiring electronics/electrical knowledge. Given that it is also known that youth are taking fewer formal courses in this area prior to entry into the military, the DoD is interested in improving its ability to select and classify recruits into electronics-related occupations. To that end, there is research in progress to improve the content of the current enlistment test. A number of large-scale research projects, on both new paper-and-pencil and computerized tests, are underway in hopes of finding better predictors of performance in military training and occupations.

The Department reiterates, however, that it is inappropriate to equate performance on the electronics composite with recruits' overall "technological sophistication" and to conclude that this sophistication has declined over the decade of the 1980s. Unfortunately, there is no way to conduct a historical study on this subject. The DoD concurs with GAO researchers that the youth and entry-level labor force demographics are changing and that the Department needs to study carefully the effects of its enlistment test and concomitant composites on the people (e.g., women, minorities) that will be recruited in the future. To that end, the results from enlistment test research described above are expected to be helpful in making future enlistment test decisions.

FINDING D: Schoolhouse Measures of Training Effectiveness--Army. The GAO reviewed course grades in Army advanced individual training courses for five occupational specialties to determine the extent to which appropriate data were available to the Military Services for use in judging training effectiveness. The GAO found that the course grades for the five specialties were not equally reliable indicators of performance during training. The GAO noted, for instance, that at Fort Gordon it was unable to

find a consistent relationship between milestone measures and final grades, nor was it able to locate anyone who could suggest a relationship. The GAO concluded that the grades recorded for two of the courses (36L and 39B) could not be used to discriminate reliably among the performance of individual trainees. The GAO found inconsistencies in scoring between different classes and even within the same class. The GAO also found that Fort Gordon's grades (unlike Redstone's grades) were based partially on measures of physical conditioning that appeared to be unrelated to job performance. The GAO concluded that the psychometric differences it found at Fort Gordon appeared to be the result of a number of factors including (1) questionable data entry procedures and software and (2) the pass/fail nature of the criteria used to evaluate student progress. GAO suggested that subject matter experts need to develop more finely tuned, objective, and reliable measures of performance than "go/no-go." The GAO noted that, because of the problems encountered at Fort Gordon, it excluded those courses from its sample of Army trainees, resulting in the inclusion of all recruits who completed 24J and 27N training between October 1987 and July 1989, and approximately one-third of those who completed 29V training during the same period.

The GAO found that, on the Armed Forces Qualification Test and the Electronic Composite, male trainees scored significantly higher than did females and white trainees performed better than minority students. The GAO further found that the training performance differences correspond with the test score differences on both tests for the racial groupings. The GAO noted that for gender, training performance differences between males and females were larger than test score differences. The GAO also found that the Electronics Composite is a better predictor of success than the Armed Forces Qualification Test.

The GAO further found that, for its entire sample, the score on the Electronics Composite explains 18 percent of the variation in course grades, more than the Armed Forces Qualification Test--and a GAO-developed "factor score," which is the weighted sum of all Armed Services Vocational Aptitude Battery subtests. The GAO concluded that, for males, the Electronic Composite score appears to be a better predictor of future performance than the Armed Forces Qualification Test. The GAO found, however, that for females, the Armed Services Vocational Aptitude Battery "factor scores" are better predictors of schoolhouse performance than the Armed Forces Qualification Test, which is a better predictor than the electronics composites. The GAO noted that for minority soldiers, the ability to predict training course grades based on test scores is the weakest of all groups. The GAO concluded that the Armed Forces Qualification Test, or some other general score form the Armed Services Vocational Aptitude Battery, may provide

a better predictor of success for women recruits in electronics-related training than does the Electronics score. The GAO further concluded that better predictors of training performance are needed for minority students. (pp. 3-1 to 3-7/GAO Draft Report)

DoD Response: Partially concur. The Army's testing procedures for soldiers undergoing Advanced Individual Training are designed to ensure that soldiers achieve specified training objectives. To accomplish this, criterion-referenced hands-on performance tests are administered and scored on a "go/no-go" basis. Such tests are routinely used in the military to evaluate training effectiveness because they provide meaningful information to course managers on student performance, as well as information on the degree to which the course is meeting its stated objectives. Given that such tests are not designed to measure the relative performance of individuals (i.e., these measures are not norm-referenced), it is neither surprising nor particularly disturbing that the GAO found such test results unsuitable for correlational analysis. Criterion-referenced measurement, such as the "go/no-go" measures used by the Army, are a psychometrically sound method when mastery learning is the goal of instruction as is the case under discussion.

As with other findings in the report that describe trends in the Armed Forces Qualification Test scores and examine differences for groups (e.g., gender and race/ethnicity), the statements about training performance differences appear reasonable. However, there are problems with some of the specific analyses the GAO indicates were performed to reach those conclusions. For example, in the Army sample, students from three courses were pooled to increase the sample size and the course grades for the various specialties were assumed to be on the same score scale, or to have the same meaning. In fact, course grades tend to be on course-unique metrics and there is no way to evaluate whether a score of, say, 90 in one course means the same in terms of competence as a score of 90 in another course. Thus, the mean reported as an average of grades for the three Army courses is not meaningful and the relationship to scores from the Armed Services Vocational Aptitude Battery is tenuous. Note that for large samples, such as white males, the differences in the score scales tend to average out and the correlation coefficients are reasonably interpretable. For small samples, however, the different scales for course grades are likely to distort the correlation coefficients and means. Since the same analyses of schoolhouse measures of effectiveness were used for each Service (Findings D, E, and F), additional comments applicable to all appear in the DoD response to Finding G, the summary finding on schoolhouse measures.

FINDING E: Schoolhouse Measures of Training Effectiveness--Navy. The GAO reported that it examined scores on four training courses-(1) Sonar Technician Anti-Sub Warfare Surface, (2) Sonar Technician Anti-Sub Warfare Subsurface, (3) Aviation Fire Control Technician, and (4) Aviation Anti-Sub Warfare Technician. The GAO found the following:

- male recruits entered training with significantly lower Armed Forces Qualifications Test scores and significantly higher electronics scores than females;
- final grades for males were slightly, but significantly lower than their female classmates, suggesting that a substantial advantage in the Armed Forces Qualification Test can overcome an advantage in Electronics; and
- minority students began training with substantially lower scores on both composites but their final grades were not significantly different.

The GAO drew the following conclusions:

- that the Armed Forces Qualification Test may be more important for training success than Electronic's;
- that for most Navy groupings, the Armed Forces Qualification Test scores are better predictors of schoolhouse performance than Electronic scores;
- that for females, the Electronics composite is the weakest predictor and the "factor score" is the strongest; and
- that the ability of any of the three scores to predict training success is weakest for minorities. (pp. 3-7 to 3-8/GAO Draft Report)

DoD Response: Partially concur. While the GAO concluded that the Armed Forces Qualification Test may be more important for predicting training success than the Electronics composite and that for most Navy groupings, the Armed Forces Qualification Test scores are better predictors of schoolhouse performance than Electronics scores, a recent Navy Personnel Research and Development Center validation report found the opposite result, with an average validity coefficient of .59 for predicting "A" school success from the Composite vs. an average coefficient of .46 for prediction from the Armed Forces Qualification Test.

The report also states that the Electronics Composite is the weakest predictor and the Factor score is the strongest for females. However, statistical results from such a small sample (76 females) would not be stable enough to warrant policy changes. The results reported by the GAO, in all probability, would not be replicated given a larger sample. Also, the adjusted validity coefficients for range restriction in report Table 3.6 show for the Female Factor Score composite an increase of .42. That result is suspect, as normally such adjustments rarely provide an increase of more than .20.

It should also be noted that only one of the four training courses represented is even open to women (Aviation Anti-Submarine Warfare Technician), which is not evident without close study of report Table 3.6. The data for males in report Table 3.6 is the result of merging four training courses and produces an unorthodox analysis that requires an explanation of grading differences which may exist for the different schools.

As with the previous finding, trends in the Armed Forces Qualification Test scores and the Electronics Composite in Navy courses, including differences for groups (e.g., gender and race/ethnicity), appear reasonable with respect to schoolhouse measures of training effectiveness. However, the problems with some of the specific analyses the GAO indicates were performed to reach those conclusions remain a factor. In the Navy sample, students from four courses were pooled to increase sample size and the assumption that course grades for the various courses have the same meaning is tenuous. That limits the confidence in interpretation of the relationship to scores from the Armed Services Vocational Aptitude Battery. Note that for large samples, such as white males, the differences in the score scales tend to average out, and the correlation coefficients are reasonably interpretable. For small samples, however, the different scales for course grades are likely to distort the correlation coefficients and means. Additional comments applicable to all appear in the DoD response to Finding G, the summary finding on schoolhouse measures.

FINDING F: Schoolhouse Measures of Training Effectiveness--Air Force. The GAO reported that it examined four Air Force courses--(1) Aircraft Control and Warning Radar Specialist, (2) Automatic Tracking Radar Specialist, (3) Photo-Sensors Maintenance Specialist, Tactical Reconnaissance Sensors, and (4) Photo-Sensors Maintenance Specialist, Reconnaissance Electro-Optical Sensors. The GAO found that, like the Navy, (1) "factor scores" are as good or better predictors than composites, (2) for the female students, the Armed Forces Qualifications Test scores and factor scores out predict Electronic scores, and (3) it is most difficult to predict course grades for minority students,

although factor scores explained 10 percent (46 percent after adjustment). The GAO concluded that because of problems with some Army data, and the special preparation of data by the Navy and Air Force, it would not be appropriate to make inter-Service comparisons or make firm judge- ments about the immediate availability of psychometrically suitable measures from the Navy and the Air Force (pp. 3-8 to 3-10/GAO Draft Report).

DoD Response: Partially concur. As with other findings in the report, which describe trends in the Armed Forces Qualification Test scores and examine differences for groups (i.e., gender and race/ethnicity), the statements about training performance differences appear reasonable. The problems with some of the analyses the GAO indicates were performed to reach those conclusions restrict interpretability of the findings, as was stated in the DoD response to Findings D and E. Additional comments appear in the DoD response to Finding G, the summary finding on schoolhouse measures. The DoD does concur, however, with the final statement in Finding F, which indicates it would not be appropriate to make inter-Service comparisons. In addition, research performed by the Air Force Human Resources Laboratory confirms many of the GAO findings about general ability (such as is measured in the Factor Scores the GAO examined) as a valuable predictor of schoolhouse performance.

FINDING G: Schoolhouse Measures of Training Effectiveness--Summary. The GAO questioned the differential success in training for males and females and for whites and minorities--and about the differential predictive validity of the Armed Services Vocational Aptitude Battery for these groups. The GAO concluded that its analysis of gender and race-related differences in mean Armed Services Vocational Aptitude Battery scores and course grades in the Army suggest that the Electronic composite is an efficient simple predictor of training success. The GAO found, however, that in the Navy and Air Force, a more complex relationship exists between the Armed Services Vocational Aptitude Battery scores and course grades. The GAO noted that gender and race-related differences in course grades were quite small compared with significant differences in Electronics scores. The GAO concluded that an advantage in more general aptitude, measured by the Armed Forces Qualification Test, can compensate for a deficit in Electronics--when the deficit is not too great.

The GAO also noted that, while the Armed Services Vocational Aptitude Battery's Electronics composite score demonstrated a moderate ability to predict training success for white students and males, it was less successful for female or minority students. The GAO concluded the Factor Score that it derived was,

in most cases, the best predictor of training success because it utilized information from all ten Armed Services Vocational Aptitude Battery subtests.

The GAO concluded that, based on its work, it was impossible to determine whether the Armed Services Vocational Aptitude Battery is a weaker measure of ability for some groups--or if some other factor in schoolhouse training contributes differentially to the success of the different groups. The GAO noted that the relative inconsistency between school grades and test scores exists and should be addressed by both the recruiting and training communities. The GAO further concluded that it will become increasingly incumbent on the Services (1) to optimize selection criteria for advanced individual technical training for women and minority groups, (2) to provide compensatory training where needed, and (3) to assure that no extraneous factors within the training environment interfere with the full achievement potential. (pp. 3-10 to 3-13/GAO Draft Report)

DoD Response: Partially concur. With respect to GAO findings describing trends in the Armed Forces Qualification Test scores and the Electronics Composite and examining differences for groups (i.e., gender and race/ethnicity), the statements about training performance differences appear reasonable. The analyses of the relationships of scores from the Armed Services Vocational Aptitude Battery (Armed Forces Qualification Test, Electronics Composite, and Factor Score) and school grades are flawed and, consequently, interpretation of the results of those analyses is doubtful. Because the same analytic procedures were used for all Services and similar conclusions drawn, the following comments pertain to Findings D, E, F, and G alike.

Problems with the analyses arise from the following sources:

- pooling students from several courses, when the grades for different courses generally are not comparable;
- correction for restriction of range on the Factor Score, which resulted in correlation coefficients that are not plausible;
- lack of regression analyses; and
- small sample sizes for females.

In each Service, students for several courses were pooled to increase sample size and the course grades for the various courses within each Service were assumed to be on the same score scale, or to have the same meaning. In fact, course grades are not normally interpretable from course-to-course, because of

between-course differences in scales and the level of competency inferred by a particular score. There is no way to evaluate whether a score of, say, 90 in one course means the same as a score of 90 in another course. (For the Army, three courses were combined, four courses for the Navy, and four for the Air Force.) Thus, the mean grades reported for courses in each Service are somewhat arbitrary numbers and their relationship to scores from the Armed Services Vocational Aptitude Battery is tenuous. Note that for large samples, such as white males, the differences in the score scales tend to average out, and the correlation coefficients are reasonably interpretable. For small samples, however, the different scales for course grades are likely to distort the correlation coefficients and means.

The correlation coefficients for the Factor Scores are suspiciously high, especially after correction for restriction of range. The Factor Scores are based on the first principal component of the Armed Services Vocational Aptitude Battery and the weights tend to be uniform (from .10 to .14). The Factor Score is the sum of the 10 subtest standard scores and the correlation coefficient could be computed using the correlation of sums. An important point is that the weights are not regression weights computed to maximize the correlation between the aptitude test scores and course grades; instead, the correlation coefficient for the Factor Score is, in effect, the average for the 10 subtests.

In previous studies, the four subtests in the Electronics Composite (Math Knowledge, Arithmetic Reasoning, General Science, and Electronics Information) repeatedly tend to have the highest correlation with course grades in these kinds of courses. As a rule, therefore, the correlation with course grades should be higher for the Electronics Composite than for the Factor Score. Deviations from this expectation may be attributed to artifacts, such as restriction of range.

The GAO report recognizes that correlation coefficients in samples cannot be compared directly because of range restriction. Adjustments are made to compensate for differences in restriction of range. The adjusted values for the Armed Forces Qualification Test and Electronics Composite are plausible in that they are consistent with other analyses; the adjusted values for the Factor Score, however, are unduly high and they lack plausibility. The procedure used to correct for restriction of range should be based on the multivariate model, which involves complex formulae and computing routines. The simpler univariate model may have been used, which could distort the adjusted values for the Factor Score.

Comparisons are made by gender and minority status based on mean scores and correlation coefficients. Conclusions about the appropriateness of the Armed Services Vocational Aptitude Battery for females and racial/ethnic minorities are then based on these comparisons. Such comparisons are a good place to start, but analyses of gender and race differences should include a comparison of the respective regression lines (slopes and intercepts), errors of estimate, and cutoff scores. Analyses of differences in mean performance on predictors, final school grades, and differences in validity coefficients are not, by themselves, sufficient. With the more thorough regression analysis, meaningful conclusions can be made about the appropriateness of aptitude tests for female and racial/ethnic minorities compared to white males.

Even if the DoD were to fully concur with the statistical analyses performed, interpretation of the results for females would remain problematic because of the small sample sizes. The number of females with course grades in the samples are 18 for the Army, 71 for the Navy, and 98 for the Air Force. With such sample sizes, differences in scales for course grades may be exacerbated; correction for range restriction could lead to illogical correlation coefficients; and regression equations with up to 10 predictor variables would result in unduly high correlation. Issues of generalizing to other samples and of making policy decisions about selecting females and assigning them to technical specialties should always be considered extremely carefully and be based on thorough analysis. Replication of results is the sine qua non of analysis and an adequate sample size is a good foundation for replication. The conclusion "that the Services should consider developing a more general ASVAB (sic) derivative such as our Factor Score to assign women and minorities to technical training" (p. 5-2 and 3) is reasonable, and could be pursued by the military manpower research community. The report provides a stimulus to continue efforts to improve the effectiveness of selecting and classifying recruits, especially for minorities.

FINDING E: Field Measures of Training Effectiveness--Army. The GAO reported that, although it was aware of numerous post-training evaluation activities performed by the individual services, only the Army could provide individual performance measures. The GAO reported that, by Army regulation, a soldier's occupational specialty performance is tested within 6 months of completion of training and every year, thereafter, under the Skills Qualification Test program. The GAO found the following regarding the Skills Qualification Test scores:

- the best predictor of Skill Test scores are final schoolhouse grades;

- the Armed Forces Qualification Test and Electronics scores were also significantly related to the Skill Test scores for whites and males, but factor scores consistently out predicted the composites;

- for females and non-white soldiers, the Armed services Vocational Aptitude Battery scores were not positively related to future performance, as measured by Skill Qualification Test scores; and

- the grades scored by females at the schoolhouse were inversely correlated with the Skill Qualification Test scores."

The GAO concluded that the traditional Armed Services Vocational Aptitude Battery scores may not be the best predictor of performance for the non-traditional soldier--that is, the female or minority, soldier. The GAO observed that better predictors of success for these groups should be found. (pp. 4-1 to 4-5/GAO Draft Report)

DoD Response: Partially concur. The GAO appears to have incorrectly assumed that Skill Qualification Tests have a common metric across different specialties, skill levels, and years. Due to the requirement to develop new tests each year, individual tests are fielded with a minimum of pretesting. As a result, means and standard deviations across a specialty and even across years within the same specialty and skill level may vary greatly. For example, in the five specialties studied by the GAO, the means on the individual skill level 1 test during 1985-1989 ranged from 74.5 to 88.4, while standard deviation ranged from 3.5 to 14.7.

During the years 1985-1989, more than 3800 different tests were administered in more than 200 specialties annually across skill levels 1 to 4. The Army Research Institute is currently analyzing this data (more than 1 million scores) and intends to report Armed Services Vocational Aptitude Battery validities by both race and gender as well as for sample size whenever sample size is adequate for such analyses. Noting the GAO concern relating to low validity for blacks and females in their study, the Army has computed validities for these groups for the 1988 Skill Qualification Tests. For 71 skill level 1 samples comprised of at least 50 females, the median corrected validity is .58, for samples of 50 or more blacks the median validity is .47; the median validity for 205 total samples is .57. While the Army understands the GAO focused only on highly technical specialties,

total accessions in the five GAO selected specialties numbered only 310 compared to more than 120,000 for all specialties during 1988.

It is suspected that the finding is affected by the small samples of females and minorities in the GAO analyses. The finding that Armed Services Vocational Aptitude Battery scores were not positively related to Skill Qualification Test scores for females and non-white soldiers is contrary to the body of research evidence for predicting training grades in the schoolhouse. The consistent finding in all Services is that aptitude scores are about equally valid for females, racial/ethnic minorities, and white males, although there may be some over or underprediction for females and minorities. Research results also show that aptitude tests predict supervisors' ratings of job performance for blacks about as well as for whites. The results presented by the GAO should be evaluated in larger samples.

The same problems noted earlier with analysis of schoolhouse training grades apply to this analysis of Skill Qualification Test scores:

- pooling of specialties--Skill Qualification Test scores are not on a common metric across specialties, and the same numerical value in different tests does not, as a rule, mean the same level of competence;
- the correction for restriction of range on the Factor Score leads to distortion in the results;
- a regression analysis is appropriate and was not performed; and
- the sample size of females (18 or 21) is inadequate to draw meaningful conclusions.

Research in progress pertaining to enlistment test development, including computerized tests, will examine implications for gender and minority subgroups.

FINDING I: Field Measures of Training Effectiveness--Navy. The GAO reported that it considered two possible sources of field information routinely collected by the Navy as measures of the effectiveness of the training courses--(1) Level II surveys and (2) Advancement in Rating Examinations. The GAO found, however, that the Level II surveys have been effectively abandoned by the Navy, with none having been performed since at least 1986. The GAO concurred with the judgement of the test developers and administrators that, because the test is not standardized and is

not administered to all graduates, the Advancement in Rating Examination is "not a good source of training evaluation feedback."

The GAO reported that, in 1986, the Chief of Naval Operations requested that the Naval Training Systems Center determine the current status of Navy training evaluation and provide recommendations. The GAO further reported that, while numerous non-formal or non-centralized activities were identified, the Naval Training Systems Center found that:

- the quality of current Navy schoolhouse training could not be readily ascertained for the vast majority of the courses being offered;
- there is a lack of technical evaluation/assessment skills; and
- current evaluation activities are fractionated, not comprehensive, and operating in an environment of obsolete instructions and unclear objectives.

The GAO reported that the Navy made a number of recommendations to upgrade and take a systematic approach to training evaluation. According to the GAO, the Navy has assigned a three-person team to review the proposals and recommend an integrated training appraisal program. The GAO concluded that, while the Navy should be commended for its willingness to acknowledge past evaluation deficiencies, it seriously questioned whether this response is appropriate to the severity and extensiveness of the problems that the Naval Training Systems Center has documented. (pp. 4-5 to 4-8/ GAO Draft Report)

DoD Response: Partially concur. Level II surveys were discontinued by the Navy because they were paper-intensive and placed an undue burden on the fleet. Moreover, only limited methods of evaluating the effectiveness of schoolhouse training were in effect at the time the Navy requested the Naval Training Systems Center to determine the status of evaluation procedures and make appropriate recommendations. Since that time, however, the Navy has successfully employed several means of collecting feedback on training effectiveness. In addition to the steps being taken by the Navy to enhance training evaluation methods as reported by the GAO, several other programs are underway. These include the (1) Navy Training Appraisal Program, (2) Navy Training Requirements Review, (3) Fleet Training Appraisal Program, and (4) Maintenance Training Improvement Program. These are discussed in more detail in the following paragraphs.

A Navy training appraisal program was implemented in March 1989. The process provides the Chief of Naval Operations with an assessment of the adequacy of Navy training to support warfighting capabilities in each of the Navy's primary mission areas and focuses attention on specific areas where training may be deficient. The training appraisal program allows scarce training assessment resources to be brought to bear upon those training programs that fleet feedback reveals are most in need of attention. The Navy training appraisal process has thus far examined acoustic operator, damage control/firefighting, electronic warfare operator/maintainer, and "over-the-horizon" targeting systems training.

There is also an ongoing Navy Training Requirements Review, which provides direct feedback between warfare sponsors, Systems Commands, the fleet, and the Naval Education and Training Command on a scheduled basis. That program requires fleet experts to talk directly to school personnel and provides valuable information on training effectiveness.

Additional training effectiveness feedback systems in place include the Fleet Training Appraisal Program and the Maintenance Training Improvement Program which provide fleet performance data. The Training Performance Evaluation Board Training Evaluation and Assessment Division was staffed in February of 1990 and has as part of its charter the study of training feedback systems.

FINDING J: **Field Measures of Training Effectiveness--Air Force**. The GAO reported that it considered sources of individual level data for field performance of Air Force personnel equivalent to those it used for the Navy, but concluded that neither the promotion examinations nor the supervisory surveys were appropriate. The GAO further concluded no individual data exist that would allow an analysis equivalent to those performed by the Army with the Skill Qualification Test data.

The GAO reported that other Air Force training assessment procedures exist, including Training Quality Reports, Utilization and Training Workshops, and Occupational Survey Reports. According to the GAO, the Training Quality Reports are part of a reactive evaluation process, while the other activities are more concerned with front-end analysis. (pp. 4-8 to 4-10/GAO Draft Report)

DoD Response: Partially concur. The Air Force is aware of the potential shortcomings of promotion examinations and supervisory surveys for evaluating training effectiveness, and is currently developing career field training management guidelines to track and enhance the training from enlistment throughout an individual's career. Emphasis will be placed on criterion-referenced

objectives rather than the present code levels for performance standards. These changes will have a major impact on the present promotion system. To expedite feedback from supervisors concerning any problems with recent graduates, a new policy was recently established by the Air Training Command to provide telephonic communication on a 24-hour basis between the training center providing the training and the supervisor of the graduate. The system allows more effective and timely communication between the supervisor and the training provider.

The Air Force does not have Skill Qualification Tests for performance and does not plan to have them in the near future. Many of the tasks performed in the field are very complex. Testing, recording, and documenting individual performance for statistics is very time consuming, requires additional manpower, and is cost-prohibitive. Further, many of the new Air Force systems are single channel systems, which cannot be used for extensive training or evaluating trainees. All these factors combine to make the use of hands-on Skill Qualification Tests an inappropriate solution to the problem of training effectiveness evaluations. The GAO finding that Occupational Survey Reports are concerned with front-end analysis is true, but information about what first-termers are doing on-the-job provides a good basis for what should be trained and what is expected in the initial skills courses. As written in the report, the paragraph gives a very limited view of what Occupational Survey Reports provide the training community and their potential for training assessment.

FINDING K: Alternative Data Sources: The Job Performance Measurement Project. The GAO reported a key impediment to establishing a field evaluation component of training assessment is the expense of developing, testing, and administering measures that validly and reliability measure actual performance. The GAO noted that, beginning in the early eighties, a major effort, entitled--"The Joint-Service Job Performance Measurement Project," designed to address the measurement issues, has been underway under the direction of the Office of Accession Policy located in the Office of the Assistant Secretary of Defense (Force Management and Personnel). The GAO reported that this project was initiated after the Armed Services Vocational Aptitude Battery unintentionally allowed some 300,000 less qualified recruits into the Military Services and resulted in field commanders' complaints of quality degradation among their personnel.

The GAO found that the Joint Performance Measurement project:

- did not set out to establish a link between schoolhouse performance and field performance;

- concluded suitable measures of field performance did not exist and undertook to develop them;
- has not reported any analyses of sex- and race-related differences, and has not addressed the schoolhouse/field connection; and
- concluded performance measures were expensive to develop and frequently costly to administer and, therefore, may not be suited to more routine use as measures of training effectiveness.

The GAO concluded that the investment made to develop the performance measures and their surrogates could prove to be more profitable if some of the measures developed and the lessons learned were more widely applied to the development of realistic assessment procedures for training. The GAO further concluded that the lack of other objective, systematically collected field evaluation data renders meaningful evaluation of training effectiveness impossible. The GAO observed that decision makers in the Congress, the DoD, or the Services can only react to problems in the field after they have become apparent and have been identified as training-related. The GAO concluded that, given the cost and complexity of today's military equipment, it is difficult to understand the lack of evaluative data to monitor how well Service personnel are being prepared to use and maintain those weapons. Overall, the GAO concluded that, among the most serious deficiencies it identified, was the inability of the Air Force and the Navy to found their evaluation of their selection procedures and schoolhouse training in systematically collected, objective field performance data. The GAO further concluded that, without good performance measurement data, the Services are not able to maximize training effectiveness, or even estimate realistically the success of their training investment in producing skilled operators and maintainers of today's and tomorrow's sophisticated weaponry. (pp. 4-10 to 5-4/GAO Draft Report)

DoD Response: Partially concur. The GAO analysis of the background, purposes, and findings thus far from the Joint-Service Job Performance Measurement Program are generally accurate. The GAO has also correctly identified that hands-on performance measures are resource-intensive in terms of labor, cost, time, and equipment, which limits their value for routine use as field measures of training effectiveness. The issue of applying job performance measurement technology to training was investigated in May 1985, when the Assistant Secretary of Defense (Manpower, Installations, & Logistics) solicited Service responses to an inquiry from Congressman Les Aspin, Chairman of the House Committee on Armed Services. One of the Chairman's questions specifically asked about Service plans for applying job performance data to training course design and evaluation. The Service responses

suggested how they anticipated potential applications of job performance measurement data. Each of the Services offered a plan for institutionalization of job performance measures and they identified training evaluation as a likely additional application of Job Performance Measurement technology, to include introducing performance measurement into the training feedback system. The resource factors identified by the GAO, coupled with the need to wait until completion of the enlistment standards setting portion of the Job Performance Measurement research, resulted in the decision to defer full-scale implementation of routine job performance data collection for all occupations.

It should be noted there is Service work ongoing that examines the link between schoolhouse performance and field performance. For example, the Army's Selection and Classification research program (which incorporates the Army's contribution to the Joint-Service Job Performance Measurement Project) is examining the link between schoolhouse performance and job performance. Schoolhouse (end-of-training) and job performance measures have been developed and administered to a longitudinal sample in several military occupational specialties. In addition, school grades and Skill Qualification Test scores have been obtained for the sample and analyses are underway. The Air Force, Navy, and Marine Corps have been performing similar analyses and the results will be applicable to understanding the link between schoolhouse performance and on-the-job performance.

Work is also underway in all of the Services to determine the efficacy of performance surrogates for specific purposes. There are technical and policy differences related to measuring job performance for validating a test and measuring job performance for evaluating a training system. Nevertheless, if research efforts are successful, it may be possible to use surrogates to develop cost-effective field performance feedback procedures that could help guide curriculum development.

RECOMMENDATIONS

RECOMMENDATION 1: The GAO recommended that the Assistant Secretary of Defense (Force Management and Personnel) direct the personnel research it coordinates among the individual Services to investigate more sensitive predictors of schoolhouse performance for women and minority students from the Armed Services Vocational Aptitude Battery data it already possesses. (p. 5-4/GAO Draft Report)

DoD Response: Concur. The Office of the Assistant Secretary of Defense (Force Management and Personnel) will prepare a memorandum to the Defense Manpower Data Center and the Services requesting that the recommended analyses be performed. We will also ensure that research in progress pertaining to computerized enlistment test development will include analyses to determine the sensitivity of the tests as predictors of schoolhouse performance for gender and minority subgroups.

RECOMMENDATION 2: The GAO recommended that the Secretary of the Army direct the Training and Doctrine Command to review the schoolhouse grading procedures identified within the report as deficient for their accuracy, appropriateness, and reliability. (p. 5-4/GAO Draft Report)

DoD Response: Concur. The Secretary of the Army will direct the Training and Doctrine Command to review the appropriateness of Fort Gordon's testing procedures and their compliance with Army policy. A plan of action to remedy any existing deficiencies will be prepared by August 1990.

RECOMMENDATION 3: The GAO recommended that the Secretary of the Navy establish a firm deadline for developing a training evaluation program and that he direct that the adequacy of current resources allocated to this effort be reexamined. (p. 5-4/GAO Draft Report)

DoD Response: Concur. The Navy has several training evaluation programs already in place. As mentioned previously, these include the Navy Training Appraisal, the Navy Training Requirements Review, the Fleet Training Appraisal Program, the Maintenance Training Improvement Program and the Training Performance Evaluation Board. Additionally, the Chief of Naval Education and Training plans to brief, by July 1990, an enhanced integrated training feedback system to the Chief of Naval Personnel. A Plan of Action and Milestones will be prepared by August of 1990 to implement that system.

RECOMMENDATION 4: The GAO recommended that the Assistant Secretary of Defense (Force Management and Personnel) review alternative measures of field performance already developed by the Services under the Job Performance Measurement project for potential applicability to training and on-the-job performance evaluation. (pp. 5-4 and 5-5/GAO Draft Report)

DoD Response: Concur. During the mid-1980s, the DoD explored applications of the measures developed in the Joint-Service Job Performance Measurement Program to training. While the decision made following that review was to defer full-scale implementation because of cost factors and the fact that techniques for develop-

ing the performance measures were still being refined, the Department will again explore the feasibility of expanding their use through the auspices of the Joint-Service Job Performance Measurement Working Group. The review is expected to be completed following final performance measurement development during Fiscal Year 1991.

Appendix VI

Major Contributors to This Report

Program Evaluation and Methodology Division	Michael J. Wargo, Issue Area Director Richard T. Barnes, Assistant Director Robert E. White, Project Manager Kurt R. Kroemer, Project Staff